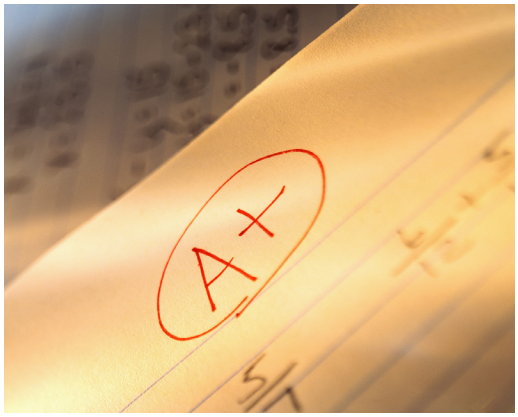
The Reliability Programme
Final Report of the Technical Advisory Group



Jo-Anne Baird
Anton Béguin
Paul Black
Alastair Pollitt
Gordon Stanley

Ofqual/11/4825

This report has been commissioned by the Office of Qualifications and Examinations Regulation.

Contents

Tables and Figures

Executive Summary

"Despite its being relatively easy to compute, there are few subjects in the field of measurement more difficult to understand than reliability"

(Cunningham, 1986, p101)

This is the Report of the Technical Advisory Group of Ofqual's Reliability Programme. It summarises the work of the programme and makes recommendations to Ofqual regarding future research and regulation. First, we note that Ofqual needs to define the purpose of the qualifications to fulfil its regulatory remit. The regulator needs to be able to address the question – *of what* are our assessments reliable indicators? This question is clearly related to validity of assessments, but there is also a reliability dimension to this issue.

There are two main reasons for regulation of reliability. The first is to foster confidence in the assessment system in England amongst users of the system. The Technical Advisory Group supports the regulation of reliability, but cautions that a judicious balance needs to be struck between publishing reliability statistics and maintaining public confidence in the system. Less than perfect reliability is a fact of any assessment system, but public understanding and tolerance of this fact might be thwarted by negative media coverage. Imperfect examinations do not create good news stories. A second, and related reason, is engendering quality improvement in the assessment system. An important role for Ofqual should be to raise standards in the assessment industry and regulation of reliability should help to do that.

Notwithstanding the above, neither Ofqual nor awarding organisations are in a position to control reliability of public assessments in England. Systemically, a number of agencies (including the Department for Education) are involved in decisions regarding the design of public examinations and assessments in England. Decisions by different bodies impact upon the reliability of the resulting assessments.

Different methods are used in the research literature on reliability, based upon different approaches to operationalising 'true scores'. Classical test theory and generalisability theory assume that a candidate's true score is the one found by averaging over many replications of assessment. For example, many examiners' marks could be averaged to find the true score. Item response theory assumes that there is an underlying, 'latent' trait being measured by assessments and scores on that trait are the 'true score'. There is no 'eye of God' reality to which we can refer to find out a candidate's true score: the score that they really deserve. Assessment scores are the product of social agreements between examiners about the extent to which different answers to questions are valued. Professional judgment is at the heart of this process and a certain amount of variability is essential if assessments are to remain valid. That is, although it would be practically possible to iron out much of the unwanted variation in assessment results, the impact of this upon assessment design and its backwash upon our education system would be disastrous. Machine-marked multiple choice tests have high inter-rater reliability, for example, but their validity for the assessment of many aspects of the curriculum is weak.

A considerable volume of data was generated on the programme with regard to the reliability of assessments in England. Nonetheless, we do not yet know enough about the curriculum and assessment format factors affecting reliability statistics. Given that we wish to retain a variety

of assessment types for validity reasons, standards of reliability expected for these assessment types must be contextualised by empirical data associated with those assessment formats and curriculum topics. Where assessments are found to be highly unreliable, this technical information can be used to inform the debate about what should be assessed and how.

Acknowledgements

The Technical Advisory Group would like to acknowledge the work that has gone into The Reliability Programme from the authors of the research reports (Appendix D), assessment experts who contributed to the seminar series (Appendix C), the Ofqual Policy Advisory Group who commented upon a draft of this report and to the Ofqual staff who initiated the programme and kept it on track. They are Dr Paul Newton (now Director of Cambridge Assessment Network), Dennis Opposs (Director of Standards), Dr Qingping He (Principal Researcher), Jo Taylor (Comparability Manager), and Annette Kinsella (Communications Manager).

Introduction

This is the Report of the Technical Advisory Group for the Office of Qualifications and Examinations Regulation's (Ofqual) Reliability programme, summarising the activities and findings and making recommendations for Ofqual's monitoring of reliability in England's qualifications and assessments. Experts in educational assessment from Australia, England, France, the Netherlands, Scotland, Wales and the United States have contributed to the programme. Original research was conducted and the programme has drawn upon the internationally published literature on reliability. In the two years of the programme, the Technical Advisory Group has advised Ofqual on a programme of research to take forward our knowledge of the reliability of England's assessments, including:

- technical measures of reliability of individual assessments and composite scores,
- use of different models to analyse reliability (classical test theory; IRT; generalisability theory),
- values for internal consistency and inter-rater reliability,
- values for the reliability of A-levels, GCSEs and national curriculum tests,
- how reliability of assessment results are reported internationally,
- the impact of standard-setting reliability
- teacher assessments
- work-place assessment.

The techniques have not been newly invented for this programme, but through this programme they have been drawn together in a collection of works that forms a resource for assessment specialists. A valuable impact of the programme has been to raise awareness of reliability techniques, analysis methods and findings across the educational assessment research community in England. In this way, Ofqual has shown leadership of the field and we consider this to be an important role for the regulator, as raising levels of understanding will help to produce improvements in assessment quality.

To interpret the findings on reliability, it is essential to understand the assessments and data in some detail. Details about the design of assessments and their resulting data can render certain conclusions naïve. Thus, our approach has been to work with a wider group of assessment experts to discuss and debate the results of the studies. This dialogue has made an important contribution to the quality of the programme and we recommend that Ofqual works in this way with academics, awarding bodies, and examination boards on future programmes.

Ofqual is in a regulatory relationship with the assessment industry, and there was initial scepticism from some staff of the awarding organisations about the need for the programme and concern about the possibility of ill-considered regulation that overly emphasised reliability at the expense of good assessment design. The trade-off between reliability and validity is a serious concern shared by stakeholders (see House of Commons, 2008 paragraph 56). Constructive engagement with assessment specialists through a seminar series, individual meetings, and other communications has produced better agreement about the need to address the reliability issues raised in the programme, as well as recognition that awarding bodies have quality assurance agendas of their own to pursue. On-screen marking has been a feature of educational assessment in England over the past few years and this has raised issues that require new quality assurance processes to be formulated, as well as research to

investigate the impact of these new systems. Equally, better possibilities for collecting, analysing, and presenting data are needed to improve the assessment design and scoring that arises from these on-screen marking systems.

Other concerns raised by these stakeholders are related to the way in which reliability data would be publicised by Ofqual. Whilst there was a willingness to raise standards in the industry in a constructive manner, this needs to be handled in a considered way to ensure that confidence in the qualifications is maintained. Tolerance of the hard fact of imperfect reliability in assessment was considered to be uncommon in the media, for example. We agree that imperfections in assessment will not create good news stories.

Some stakeholders raised concerns that reliability statistics might be used as a marketing tool and that there could be competition between awarding bodies on those grounds. Whilst this is possible, we do not consider this to be a successful business strategy unless there is wide variation between qualifications. In this instance, it would be a reasonable concern for teachers and therefore in the interests of the education system to publish such figures. The formal relationship between the regulator and the examining bodies caused some reluctance to produce reliability figures for the assessments, lest they result in a reason for the regulator to investigate the examining body. Furthermore, the figures could be publicised, which could result in even more problems for the examining board. During such discussions, the view taken was that it was better to use reliability statistics for quality improvement than to raise issues unnecessarily in a public forum. This raises a number of important points. The Technical Advisory Group is in favour of the collection of reliability statistics with the aim of improving the assessments, as well as industry skill and knowledge. Yet, it would be unwise to unnecessarily undermine confidence in the qualification system whilst attempting to improve the reliability. Some industries are able to operate in such a way that information can be shared between those who understand the data without it becoming subject to uninformed public debate. Under the Freedom of Information Act, this is not a possibility for any data that Ofqual collects. Therefore, it is essential that Ofqual strikes a balance between moving regulation of reliability forward and maintaining confidence in the system. Their communication strategy will be central to achieving this balance. Ofqual's Policy Advisory Group have a remit to advise upon the communication strategy for this programme.

Although the focus of many of the research papers on this programme has been general qualifications such as GCSE and A level, Ofqual's remit extends beyond general qualifications to vocational qualifications. Its regulatory function has to oversee over 5,000 qualifications offered by 175 awarding bodies. We note that the reliability programme has focused largely upon general qualifications because the body of evidence previously available and the expertise available to conduct further research was mainly associated with that sector.

This report outlines the research activities that have been undertaken and considers the outcomes of the programme in relation to its remit. The remit categorised four different factors that could impact upon reliability: occasion, tests, markers and standard-setting. Special considerations are given to students who take the test under difficult circumstances, such as following a close family bereavement (Ofqual, 2010, Section 7, p47), so the programme did not conduct much work on occasion. For many national qualifications, the examinations are taken at a fixed time and occasion factors affect individuals' scores positively and negatively. For example, some students will be in good form and others will be tired, coming down with a cold, or have just split up with their girlfriend or boyfriend. Research on specific occasion-related factors, such as the timetable of examinations being taken by different students, would be

warranted in future, but the strategic decision taken for the programme was to attempt to make progress with broad issues within the two-year time span. Indeed, three of the reports on public perceptions of reliability produced for the programme indicated that some occasion-related error is seen as the luck of the draw in our educational culture (Ipsos Mori, 2009, p19; Chamberlain, 2010, section 5.3; He, Opposs and Boyle, 2010, Figure 11).

Why reliability is important

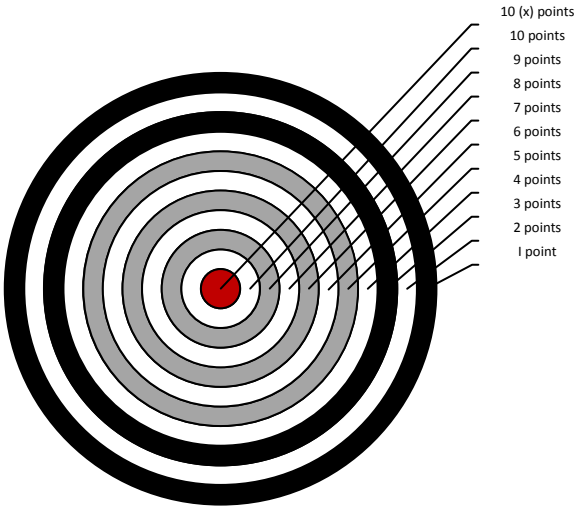
Fairgrounds sometimes have archery stands where you can pay to shoot arrows at a target and win a prize if you hit the bull's eye. As with many of these fairground competitions, they are an unreliable indication of skill largely because the bow and arrows behave erratically. Skilled archers are likely to be thwarted most of the time, yet novices might sometimes strike lucky. Historically, selecting individuals with good archery skills was an important military consideration in England. A test of archery that was unreliable would have had dire consequences for the battlefield, resulting in the wrong people in the wrong jobs. Had the fairground archery stand been used to select archers, some of the best would have been left on the bench and lucky, but terrible archers would have got their jobs. Once news got out that the archery test was so unreliable, trainee archers might be less motivated to learn their trade because it would not help them in getting through the selection process.

The fairground archery test could be improved with better archery tools so as to better reflect the on-the-job military situation. Then we would be more able to assess the underlying archery skill and could generalise our assessment scores to the real life situation because it has more validity (i.e. measures what it is supposed to measure). We could still question the validity of our improved assessment, as it does not reflect the entire real-world task. For example, the enemy does not typically stand still with a coloured target marked on his shirt, so a further improvement might be to have a moving target. With a static target, we have a standardised task, which is advantageous because it controls some of the unwanted variability that can arise in real settings. Targets are not all the same size, nor moving at the same rate in the battlefield, but allowing these things to vary in the assessment would simply give some archers advantageous conditions of larger, slower targets, and vice versa. In controlling the assessment task, there is more equitable testing, but we pay the price of this in terms of having a less authentic assessment, whose scores might not generalise so well to the real-life task.

Then there is the scoring to consider – how many more points should be given for hitting the bull's eye? Standard practice is to give one point for hitting the outside of the target, with the points increasing by one for each ring closer to the centre that the arrow hits (

Figure 1). Tie-breaks in terms of points are settled by the number of bull's eyes scored. Now, this scoring system is fine if we are interested in incrementally scoring the skill of getting closer with your shots to the centre of the target. However, if we know that the target used in the assessment is too big in comparison to real-world targets, perhaps we ought to give no points at all for hitting the outside of the target. Alternatively, we could give much higher points for hitting the target more centrally and especially for getting a bull's eye. The validity of the scoring mechanism and the test conditions are very important considerations. No matter how reliably the scoring is applied, if the test is not a valid measure of archery skill, it is not useful for selecting archers.

Figure 1 **Archery scoring**



Of course, even with perfect bows and arrows people will vary in their performances and even Victor Ruban, the Ukrainian 2008 Beijing Olympics archery gold medallist, will occasionally have a poor shot. Variability in performance is to be expected, so we need a good enough sample of people's performances to be able to draw general conclusions about their archery skill. If only one shot is allowed, unwanted variability in the form of flukes and blunders will affect our assessment.

But unwanted variability in the archery assessment could come from a variety of sources beyond flukes and blunders. On a very windy day, when the light is fading and the fairground is noisy, we might expect people to perform differently than under conditions of fair weather, bright sunlight and quietness. Therefore, the **occasion** affects scores in an unhelpful manner. The **form** of test could influence the variability in scores. The same people might perform very differently with different equipment, for example. Those who took the scores might be unreliable – their eyesight might be poor, they could be disorganised or might have illegible handwriting. Thus, inter-**rater** reliability is an issue. Additionally, there would need to be a cut-off point at which archery skill was deemed high enough to be offered a job. In other words, an archery standard would need to be set. Different **standard-setters** might have different views about what that cut-off should be. This is another source of unwanted variability.

Fortunately (for England), archers were selected in a more reliable manner than our fairground assessment. This fictitious example serves to illustrate a more general and serious point, as getting the right skills and talent in the right places is key to developing the potential talent in the country. Doing that requires a system for assessing people that gives a good indication of their capabilities. Those assessments need to give us information about the knowledge, skills and understanding that are of interest – in other words, be valid. They also need to give us that information consistently to be valid, hence the importance of reliability in assessment.

Who is responsible for reliability of educational assessments in England?

Reliability is a key part of the Ofqual remit. The *Apprenticeships, Skills, Children and Learning Bill* (House of Commons, 2009) established Ofqual, giving the organisation objectives in the following areas (paragraph 125):

(a) qualifications standards,
(b) assessments standards,
(c) public confidence,
(d) promoting awareness and benefits of qualifications
(e) efficiency.

The first three of these objectives are underpinned by reliability, with Ofqual being required to ensure that qualifications and assessments give reliable indications of knowledge, skills, understanding, achievement, and attainment (including over time). In promoting public confidence in qualifications, Ofqual are required to assure not only their reliability, but that they are trusted (paragraph 344). Under Ofqual's remit, they are required to notify the Secretary of State of significant problems in the qualifications system and lack of reliability is one of the examples of such a circumstance arising (paragraph 420). However, Ofqual does not design or administer educational assessments. Rather, their responsibility is to set out regulations and monitor the country's assessments against them. It follows that Ofqual alone cannot control the reliability of educational assessments and needs to work through and with other agencies to have an impact upon reliability.

So, if the regulator does not control reliability, we might look next to the awarding bodies and examination boards who *do* design and administer the assessments to shoulder the responsibility of producing high reliability. Again, the reality is more complex than this, as awarding bodies do not have free reign to design the assessments. For many educational assessments (such as GCSE and A-level), they must comply with the Qualifications and Curriculum Development Agency's (QCDA[1]) qualifications and subject criteria. Additionally, the design of the syllabuses and question papers is approved by Ofqual and QCDA. Fundamental issues such as the coverage of syllabus content and the format of the questions are considered in this process. For national curriculum tests, the Department for Education would have to approve any major change to the design of the tests. The picture is not quite the same for vocational qualifications, in which the form of assessment is less tightly controlled. For different qualifications, then, there is an incomplete and variable control of reliability by the awarding body or examination board. Large-scale public examinations such as GCSE are designed jointly by a number of agencies, none of which have complete control over the process. Decisions taken at different levels of the system have an impact upon the reliability of the resulting assessment, such as: curriculum decisions, number of grades, number of marks available, difficulty of the assessment for the candidature, sampling of the curriculum, predictability of the assessments, administration arrangements, examiner appointments and so on. The technical requirement for reliability might not be prioritised (wittingly or unwittingly) when decisions are taken.

[1] Note that the coalition Government has indicated that QCDA will be abolished. It is currently unclear whether responsibility for curriculum matters will be taken on by the Department for Education or another body.

A more extreme instance is the new, composite qualifications that make up the Diplomas or vocational awards through the Qualifications Credit Framework (QCF). To get the overall award for composite qualifications, students must take a combination of assessments that might be administered by different examination boards. For the Diploma, the Principal Learning, functional skills and additional and specialist learning might be offered by a number of providers. In these cases no single body could be held accountable for the reliability of the overall qualification. As such, the reliability of the overall qualification cannot be regulated because there is no vehicle by which Ofqual can regulate. The best that can be accomplished in these cases is regulation of the reliability of the composite parts of the qualification. A further issue arises in relation to regulation of the reliability of scores used for particular purposes, which are discussed further below.

The Reliability Programme

The House of Commons Children, Schools and Families Committee's report on Testing and Assessment (2008) concluded that the Government should set out clearly the purposes of national testing, alongside the evidence on validity and reliability of the tests for each of those purposes (paragraph 61). This is a view shared by the International Test Commission and the British Psychological Society (2000, p10). Additionally, the Select Committee considered that "estimates of statistical measurement error be published alongside test data and statistics … to allow users of that information to interpret it in a more informed manner" (2008, paragraph 62).

Subsequent to the House of Commons Report on Testing, Ofqual initiated a research programme on reliability of assessment in 2008, to be completed in 2010. The definition of reliability for the programme was as follows,

> Reliability refers to the consistency of outcomes that would be observed from an assessment process were it to be repeated. High reliability means that broadly the same outcomes would arise. A range of factors that exist in the assessment process can introduce unreliability into assessment results. Given the general parameters and controls that have been established for an assessment process – including test specification, administration conditions, approach to marking, linking design and so on – (un)reliability concerns the impact of the particular details that do happen to vary from one assessment to the next for whatever reason.

The remit for the work, written by Paul Newton, is given in Appendix C. The Technical Advisory Group have contributed to the first two strands of the programme , generating and interpreting evidence on reliability, by advising Ofqual on the overall shape of the research programme, on the specifications for research tenders and assisting with the quality assurance of the resulting reports.

A series of five expert seminars was held. The seminars focused upon the interpretation of the research on reliability arising from the programme and beyond (strand 2 in Appendix C). A list of experts who contributed to the seminar series can be found in Appendix B. Additionally, Ofqual held a public seminar on reliability presented by Professor Dylan Wiliam. The Technical Advisory Group have been kept informed of Ofqual's activities in relation to development of a policy on reliability (strand 3), and two members of the Technical Advisory Group (Baird and Black) were also members of the Policy Advisory Group.

The evidence submitted to the House of Commons Select Committee, as discussed earlier, exposed the level of knowledge on the reliability of public examinations. Professor Dylan Wiliam (see also Wiliam, 2001) presented the findings of a simulation (House of Commons Report on Testing and Assessment, 2008, paragraph 51), whose findings implied that there was a rate of 30% of students being given the wrong national curriculum test level result, due to unreliability in the design of the test (internal consistency). Jon Coles, a Director of the Department of Children, Schools and Families[2] was convinced that the error rate was not as high in GCSEs and A-levels, and Dr Ken Boston, Chief Executive of the Qualifications and Curriculum Authority[3] (QCA), was surprised by the figure being as high as 30% for national curriculum tests. None of the information presented to the Committee on reliability was based upon empirical research using data from the examinations, so none of the parties had a firm answer about just how unreliable our public examinations were. Professor Paul Black inquired of QCA whether they had such evidence, but the response was that there was "little research into this aspect of the examining process" (paragraph 54, House of Commons Report on Testing and Assessment, 2008). (Subsequently, Newton (2009) published a range of statistics on the reliability of national curriculum tests, but as he pointed out, the information did not answer all of the questions.) Neither had there been an equivalent publication on GCSE or A-levels, or any of the other qualifications that Ofqual regulates. The reliability programme has contributed to our understanding of the reliability of these qualifications. More specifically, we know from empirical research using national curriculum test data that the 30% figure was too high and should have been closer to 10% in mathematics, 13% in science, 14% in English for the internal consistency of Key Stage 2 tests (He, Hayes and Wiliam, 2011).

Returning to the Select Committee recommendations, it was recommended that the Government clarify the purposes of national assessments and that the reliability of scores for these purposes be published (paragraph 61). National assessments such as national curriculum tests, GCSEs, A-levels, Diplomas, BTECs and the International Baccalaureate serve multiple purposes under Newton's classification scheme (Table 1), and there has been no statement from Government which prioritises a specified use of any of the assessments over the others.

Table 1 Newton's 22 purposes of educational assessments

1) social evaluation	to judge the social or personal value of students' achievements
2) formative	to identify students' proximal learning needs, guiding subsequent teaching
3) student monitoring	to decide whether students are making sufficient progress in attainment in relation to expectations or targets; and, potentially, to allocate rewards or sanctions
4) diagnosis	to clarify the type and extent of students' learning difficulties in light of well-established criteria, for intervention
5) provision eligibility	to determine whether students meet eligibility criteria for special educational provision

[2] Now the Department for Education
[3] Subsequently the Qualifications and Curriculum Development Authority

6) screening	to identify students who differ significantly from their peers, for further assessment
7) segregation	to segregate students into homogeneous groups, on the basis of aptitudes or attainments, to make the instructional process more straightforward
8) guidance	to identify the most suitable courses, or vocations for students to pursue, given their aptitudes
9) transfer	to identify the general educational needs of students who transfer to new schools
10) placement	to locate students with respect to their position in a specified learning sequence, to identify the level of course which most closely reflects it
11) qualification	to decide whether students are sufficiently qualified for a job, course or role in life—that is, whether they are equipped to succeed in it—and whether to enrol them or to appoint them to it
12) selection	to predict which students—all of whom might, in principle, be sufficiently qualified—will be the most successful in a job, course or role in life, and to select between them
13) licensing	to provide legal evidence—the licence—of minimum competence to practice a specialist activity, to warrant stakeholder trust in the practitioner
14) certification	to provide evidence—the certificate—of higher competence to practise a specialist activity, or subset thereof, to warrant stakeholder trust in the practitioner
15) school choice	to identify the most desirable school for a child to attend
16) institution monitoring	to decide whether institutional performance—relating to individual teachers, classes or schools—is rising or falling in relation to expectations or targets; and, potentially, to allocate rewards or sanctions
17) resource allocation	to identify institutional needs and, consequently, to allocate resources
18) organisational intervention	to identify institutional failure and, consequently, to justify intervention
19) programme evaluation	to evaluate the success of educational programmes or initiatives, nationally or locally
20) system monitoring	to decide whether system performance—relating to individual regions or the nation—is rising or falling in relation to expectations o r targets; and, potentially, to allocate rewards or sanctions
21) comparability	to guide decisions on comparability of examination standards for later assessments on the basis of cohort performance in earlier ones

22)	national accounting	to 'quality adjust' education output indicators

(from Figure 1, House of Commons Select Committee Report on Testing and Assessment)

Superseding reliability is the issue of validity and the use of test scores for these multiple purposes, that is, are these assessments valid measures of for example, teaching quality? Popham (2007) questioned the sensitivity of tests designed to give feedback to students to differentiate in terms of teaching quality. As previously stated, if the measures are not valid, reliability is of no consequence. Assuming validity, reliability is an issue. Use of assessment results to measure institutional quality in the performance tables in England has been questioned by some researchers, as apparent differences between institutions largely disappear once the previous attainment of students is taken into account (Leckie and Goldstein, 2009). Raw statistics in performance tables therefore give an unreliable indicator of institutional performance because they do not account for small sample sizes within institutions within any year or the characteristics of the intake of pupils.

Equally, we could look at the capability of national curriculum tests to measure reliably a cohort's achievements in mathematics, science or English. An obvious problem is that the reported percentages of students who reach each level are not measured very precisely. When standards are set, the panel of examiners have to choose a particular mark on the test to be the cut-score for a level, with those below being awarded the lower level. Often, three percent of the cohort will have been awarded the same total mark for their performances on the test (e.g. 123 out of 150). When this is the case for a range of marks (as it often is), a change in the cut-score by one mark will affect the reported percentages by 3%, which can give the impression that standards have risen or fallen to a larger extent than examiners might have intended. These are but two examples of problems relating to reliability of differing purposes of assessment results – many more could be cited and these are large research areas in their own right.

The reliability programme prioritised research on the consistency of scores for individual students for English assessments. Aggregation of scores for other purposes (such as institution or system monitoring) raises separate and important validity and reliability issues that should also be investigated. No definition of the primary purpose of the qualifications being researched was given in this programme. Where assessments are put to inappropriate uses, the major problem is lack of validity, with reliability only a subsidiary problem. The lack of clarity over the purpose of public examinations in England leaves them open to competing expectations from stakeholders and undermines public confidence. Without clearly defining the purposes of the qualifications, it is impossible for the regulator to meet its remit. The regulator needs to be clear about which purposes it is regulating. After all, the assessments might be highly reliable and valid for some purposes, but not for others. Where more than one purpose of a qualification is identified, the compromises that need to be made to achieve multiple purposes should be clarified. For example, if an assessment is to be used for qualification and selection, it might be that different numbers of candidates are awarded the grade in different years, depending upon the numbers taking the qualification and achieving the criteria. Stakeholders would need to be made aware of this expectation.

Recommendation 1 **Ofqual should outline the primary purpose of each qualification and Ofqual should regulate against that purpose.**

The above recommendation would require negotiation and consultation with a wide range of stakeholders, including Ofqual. This recommendation is in keeping with the 2008 Select Committee recommendation and the draft *European Framework of Standards for Educational Assessment* (van Lent, Watts and Wools, 2010, section 3.2.1), which recommends that the goal of an assessment should be specified, and include:

- what the assessment measures
- what inferences can be drawn from the results
- who the intended users are
- who the intended candidates are

Assessment design

Although it is counterintuitive, it would be dumb to insist upon the highest levels of reliability for every educational assessment. Very high levels of reliability can be achieved by constraining the assessment design to multiple-choice, machine-marked tests. Those tests are suitable for some subject matters and are, for example, used frequently in medical educational assessments. However, we would not want all assessments to be designed with only reliability in mind because that strategy has the potential to undermine validity. Extended writing, musical performances, presentations, science practicals and writing IT programmes are all important parts of education that we would like to incorporate in our educational assessments, despite the knowledge that the scoring of these will be less reliable than multiple-choice tests. There is a balance to be struck between decisions on curriculum, validity, standards and other matters and the effects that the outcomes of these decisions will have upon reliability. After all, many assessment design decisions are taken with regard to the impact they will have upon the education system more generally. As discussed previously, decisions which are taken at different levels of the system have an impact upon reliability. Bearing this in mind, the question of the acceptability of a particular level of reliability becomes; how reliable does an assessment *of this design* have to be before it is considered acceptable? Of course, the possibility remains that some assessment designs might be so unreliable as to be entirely unacceptable.

Interpretation of the reliability evidence

One of the big questions for this programme has been when an assessment is reliable enough. The most recently published claim for the reliability of general qualifications in England came from the Schools Council (1980), with reliability of examination grades on a seven point grade scale being claimed to be correct to within one grade. Given the importance of educational assessments for people's life chances, assessment error is problematical. However, assessments in other aspects of life, such as interviews or medical assessments are at least as important and there is an understanding that they cannot be perfectly reliable. So it is with educational assessments, yet the question remains – how reliable is good enough? At one end of the scale, tests would be so unreliable as to be useless. The fairground archery test mentioned earlier would come into this category. Perfect reliability is theoretically possible, but there are so many possible causes of unreliability that it would be unlikely. Thus, there will be a range of reliability scores for tests within which we must set criteria for acceptability. To do that, we would have to define acceptable ways of measuring and reporting error and select particular points on a scale below which the reliability will be deemed unacceptable.

Few guidelines are available for interpreting the value of reliability evidence. The obvious reason for this is that the reliability indices will largely depend on the population for which the

test is used, and the conditions for the administration (e.g. how much time is available, what is the motivation of the respondents etc). For example, it is hard, if not impossible, to design a writing assignment that has a valid marking scheme and for which at the same time the reliability index is relatively high. The *Dutch Association of Psychologists* (Evers, Lukassen, Meijer and Sijtsma, 2009) give some general rules. For reliability of high-stakes tests the guidelines are that a test with a reliability index above 0.9 is good, between 0.8 and 0.9 reliability is scored as sufficient, and below 0.8 reliability is considered insufficient. For lower stakes test the thresholds are 0.1 lower, and for reporting on group level the threshold is 0.2 lower. Another exception is made for sub-test for which the thresholds are 0.1 lower than for the total test. So a sub-test of a high-stakes test should have a reliability index above 0.8 to be evaluated as good.

The British Psychological Society stated that reliability is context-dependent in their test review procedures and give only broad guidance (Lindley, Bartram and Kennedy, 2008, p17). Currently, the European Framework (van Lent, Watts and Wools, 2010) does not specify values for acceptable reliability. Instead, it indicates that technical reports should provide evidence of the reliability and validity of the assessments (section 3.2.2). We outline in more detail what should be contained in the content of those reports in

Figure **3**.

Although it is not currently possible to specify acceptable values for reliability for assessments regulated by Ofqual, it would be possible with more information. Standards for reliability of particular qualifications should be empirically grounded in data on reliability for assessments of different formats. We recommend that Ofqual should set out standards based upon empirical work within five years. To do that, a body of data on the reliability of a range of assessments must first be collected.

Recommendation 2 **A body of data should be collected by Ofqual on the reliability of a range of assessment types.**

We envisage that monitoring of the reliability of assessments would be contextualised by these data, such that a figure for the reliability of an extended writing examination in English might be judged to be comparable (or otherwise) with examinations of that type. As part of this programme, Wheadon and Stockford (2011) and Bramley and Dhawan (2011) published a range of reliability statistics for operational data generated from GCSE and A-level examining. Separately, Newton (2009) published a range of reliability statistics relating to the national curriculum tests and these have been augmented in this programme (Maughan, Styles, Lin and Kirkup, 2009); Hutchison and Benton, 2009; He, Hayes and Wiliam, 2011).

Questions, question papers and qualifications

Another issue that has arisen on the programme is the level at which reliability statistics should be collected and reported. Most often in the literature, statistics are reported for question papers because administration of examinations has traditionally been organised by question paper in England. Examiners would have been asked to mark entire question papers for a particular examination, separate standards are set for each question paper and marking quality assurance was typically conducted for each question paper. Statistics for reliability of question papers best serves the purpose of improvements in assessment design. Reliability statistics for the overall qualification provides information on the grading of the assessment, which affects people's subsequent occupational and educational careers. Users of the assessment are therefore likely to be most interested in reliability statistics at qualification level.

Qualification level reliability will usually be better than that of its component parts because error is random and results are aggregated over components with a compensatory aggregation method. Errors are distributed according to the bell-shaped or Gaussian curve. This was first proposed by Gauss in 1794 and was known as the 'astronomical error law' in the 19th century because of its application at that time to the study of astronomy data (Porter, 1986, p.6). A simple illustration is given in Figure 2. For each of the two tests, a student deserves 50 marks – that is her 'true' score. On the first test, an examiner gives her 55 marks and on the second, she is given 48 marks. Thus, the level of error on the tests was +5 and -2 respectively. When the marks are added together, we know in our omniscient example that the student deserves 100 marks, but she will be awarded 103, an error level of only 3 marks. In this case, the size of error for the overall score is lower than that for the two tests combined (an absolute value of 7). Errors are not correlated, so we would not expect an association between positive error (extra marks) and negative error (too few marks) for individual students. Therefore, when they are aggregated, there is less error for the population of scores. We note that it is not necessarily the case that aggregate qualification error will be lower for any *individual* student, only for the population as a whole.[4]

Figure 2 Aggregation of error

With the introduction of electronic processes in examination marking, the process is changing. Rather than an examiner marking the entire question paper, individual questions or groups of questions are often marked by an examiner. This has implications for the collection of information about reliability of individual examiners, as in such a system the statistics would relate to the questions rather than the examination paper overall. In keeping with the above, we clearly see that the reliability of individual questions and sets of questions are important, but view the grading reliability as the most important focus. Equally, researching and reporting the reliability of component parts will support progress on improving assessment design, so we wish to encourage research on component reliability.

He (2009), and Bramley and Dhawan (2010) set out the approaches and complexities of calculating qualification level reliability. There are significant issues involved in collecting and reporting qualification level reliability, particularly with modular examinations, because the

[4] This might not apply to qualification grading misclassification in certain circumstances (e.g. there is a small range of marks between the grade cut-scores).

statistics are generally calculated for a population, rather than an individual and the population who takes the same tests is diminished in modular structures.

Recommendation 3 **Where possible, reliability statistics for the qualification as a whole should be produced because information at this level is important for assessment users.**

Modelling the world – examination data and 'true' scores

Variability in assessment results is caused by legitimate *and* unwanted factors. Legitimate causes include students' ability levels, effort and good resources. Unwanted factors include those that were identified earlier as producing unreliability. The Technical Advisory Group considers that assessment design should have as one of its aims the reduction of unwanted variability: that variability which is not associated with what the test is trying to measure. Unwanted variability is a better term than error because true scores are never observed. In fact, true scores do not really exist. True scores are useful constructs to help us model the world, but that does not mean they should be reified. Lumsden (1976, cited by Saal, Downey and Lahey, 1980, p.417) refers to true-scores as 'an eye of God reality'. As we saw in the archery example earlier, how many marks are given for a shot is a matter for social agreement. If an arrow is on the boundary of two rings, there will be variability of the scores if repeated scorings are made. Calling this an error disguises the fact that the assessment decision is a professional judgment. As such, 'unwanted variability' better reflects the reality because it recognises the fact that, though we may wish it otherwise, professional judgments vary at times in an unwanted, but valid manner. Whether an essay deserves a score of 15 or 16 does not have an unequivocally correct answer. These matters are settled through discussion and by designating some views, those of the senior examiners, as more powerful than others.

Many of the programme's reports referred to the different ways of approaching reliability that follows from the three models most commonly used currently:

- Classical test theory (or true-score theory)
- Item response theory
- Generalisability theory

The meaning of true scores differs in these approaches, as do ways of calculating and representing reliability. In classical test theory, the mean of all possible markers and all possible test conditions and items is deemed the true score. For item response theorists, an underlying latent trait is being measured and the true score is on that dimension. Generalisability theory treats true scores in much the same way as classical test theory, but allows investigation of different error sources simultaneously. Latent traits and averages of all possible marks under all conditions are theoretical specifications of true scores that are not observed in reality, but are the basis for the models upon which the methods of analysis are based.

In real life, operationalisations of true scores need to be created by assessment researchers seeking to establish the reliability of a test. For those adopting classical test theory or generalisability theory, the average of marks collected over different examiners (or items or occasions, depending upon the facet of reliability under scrutiny) is often the operational definition of truth. For item response theorists, a scale is created through statistical analysis and the score on that scale is their operational definition of latent score truth. Different replications lead to different data and therefore can lead to different measures of true score for the same people sitting the same assessments.

Another approach to defining true scores is to deem the most senior examiner's marks as true. This is the traditional operational approach in general qualifications in England and it has sometimes been used in reliability research using classical test theory methods. A variation on this theme is to assign a panel of senior examiners to agree the true scores, typically for a batch of responses to be used in a study or for marking quality assurance benchmarks.

We do not have good theories about the constructs that we are trying to measure in our educational assessments, nor what might cause the scores on our tests, or the relationships between scores and the underlying construct. Borsboom (2006) makes this point more generally about psychometrics and argues that we must not fall into the trap of thinking that our operationalisations in the form of test scores are equivalent to construct scores, calling this the 'operationalist thesis'. Despite active research on learning in many subject areas, models of learning that underpin test design are either not referred to or remain 'in … puberty, infancy, or even at the fetal stage' (Sijstma, 2006, p.453). To be blunt, we are not very explicit about what our tests are trying to measure and this causes problems for clarity in quantification of whether our measures are consistent. This is a general feature of the educational assessment field and not particular to England.

As there is not a consensus in the literature about the best way of modelling assessment data, we have no rationale for imposing one of the approaches outlined above over the others. However, as outlined below, more explicit reporting of the model used would assist with this.

In the next sections, we look at the research that has been conducted on factors affecting reliability that are related to tests, markers and standard-setting. Test-related factors are covered in two parts: those internal to the test and those caused by having more than one form of the test. The programme repeatedly came across specific issues related to teacher assessment and work-place assessment. Teacher assessment is at least a component part of many qualifications and often forms the entire qualification. Little research has been conducted or published on work-place assessment. As such, each of these deserves special attention and we have devoted a section to each of those to outline the issues that arose and the research undertaken. Less progress has been made on the programme with teacher assessment and vocational assessment, although conceptually the issues are the same as for other assessment forms.

Forms of reliability – evidence, reporting and policy

A lot of the angst about reliability of examinations in England focuses upon whether the examiners (raters) have got the marking right. A range of studies have been published nationally and internationally on rater reliability and how it should be measured and reported, and these issues have also formed part of the research that has been conducted on this programme. However, seldom is there conceptual clarity in the published literature about what is conceived as reliability in these studies. This point was raised by Newton (2009) and he expanded upon it at the October 2009 Reliability Programme Seminar (see Table **2**). One issue is whether researchers believe they are comparing sets of observed scores, or observed scores with true scores (or as close to true scores as we can get operationally). If the true scores are not known, we can only draw conclusions about consistency. These are boxes 1 and 2 in Newton's table below; in box 2 he looks more widely than marking to the replication of the assessment. Theoretically, true scores need to be defined in their broadest sense, as it would be odd to talk about a true score for the marking process as distinctive from a true score taking into account replications of assessment. In practice, studies often look at marking separately from other facets.

Table 2 Rater reliability conceptualisations

Assessment property	**Descriptor**	**Comparison**	**How likely is it that students would be awarded … grades**	**Question stem completion**	
				For marking	**For all facets**
Reliability	Consistency	observed vs. observed	different	… if the marking were to be replicated? ❶	… if the assessment were to be replicated? ❷
	Correctness	observed vs. true	incorrect	… given their true performance scores? ❸	… given their true test scores? ❹
Validity	Accuracy	observed vs. true	incorrect	n.a.	… given their true construct scores? ❺

Adapted from Newton Reliability Programme 2009 presentation

Some studies have included a set of marks that the researchers can claim to be true and therefore make statements about incorrectness (box 3 in Newton's classification above). Newton argued that when we are looking at how well students are classified in relation to the grades they deserve, according to their ability on the underlying construct (box 5), we are straying into the realms of validity and beyond reliability. He reserved the term 'accuracy' for

that situation. Certainly, constructs are theoretical entities that are never observed in practice, so we do not have any definitive evidence in relation to box 5.

The problem with trying to make a general claim about the reliability of a test score is that the perfect study including all possible replications that will generalise to theoretical forms of true score is never conducted. We never replicate markers and assessments and occasions. To do so would be prohibitively resource-intensive: some of the inter-rater studies that have been conducted have cost six figure sums as it is. The point about limitations in the design of studies to investigate all possible forms of reliability contemporaneously persists no matter the methodological or theoretical stance taken in relation to reliability. This begs the question of what is intended by true score, as it makes no sense for it to be estimated separately in studies investigating different facets of reliability (e.g. markers or items).

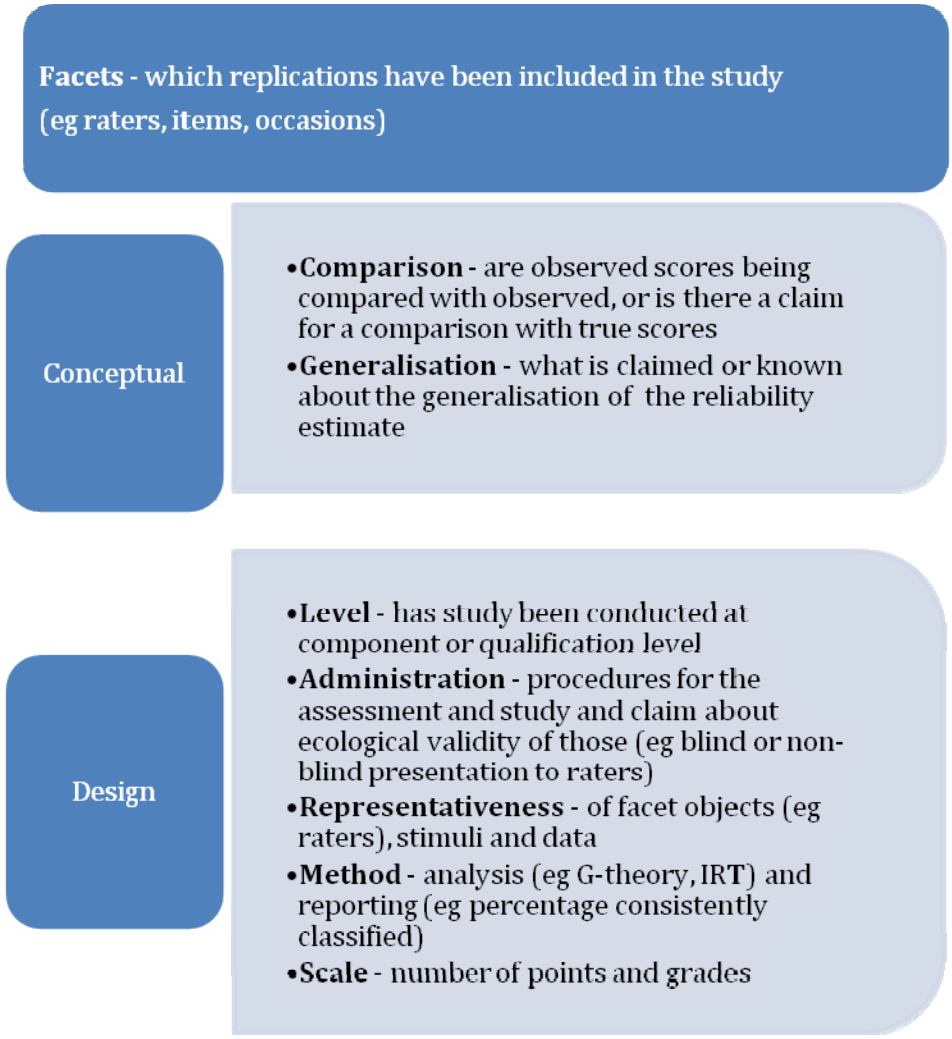
Over the past decade, there has been debate about the extent to which reliability is a feature of tests or scores; the latter being population- and administration-dependent (Thompson and Cook, 2002). A number of articles have been published using a variety of methods to investigate reliability generalisation (for example, see Special Issue of *Educational and Psychological Measurement*, 2002, p.62). 'Generalisation' in this context means the extent to which reliability statistics will pertain if the sample or some other aspect of the replication is changed. That is, if we have a misclassification figure of 20% for a test, will that rate hold if, for example, we test students in 2015 rather than 2010? Reports of a lack of generalisation have led some to use the term datametrics, rather than psychometrics, meaning that the properties of the assessment change with the data. Some authors have assimilated findings on reliability of certain assessments and conducted meta-analyses to determine the level of generalisation that is found empirically. However, Dimitrov (2002, p.792) argues that meta-analytic techniques cannot be used to investigate more than one type of replication in a reliability generalisation study. The conclusion from this area of research is that there are different facets of reliability to be delineated and the measurements of reliability might not generalise to a different population of interest. We do not know much at all about reliability generalisation or stability for English qualifications.

In the literature on validity, Cronbach (1988) moved the field forward by writing that validity is not a property that either exists or does not exist. Instead, the approach that is now generally accepted is that a validity argument must be put forward, which outlines the particular claims that are being made and the evidence for those. The above leads us to suggest that a reliability argument should be put forward when publishing findings, in which evidence is presented for reliability of scores and the claims for reliability should be clearly outlined.

Recommendation 4 **Awarding bodies should document the reliability of their assessments using the checklist for reliability claims (**Figure 3**).**

We propose a structure for documentation of reliability claims for assessments (Figure **3**) and discuss each of the content areas below. The reporting might include several studies, in which case the content should be described for each study. This proposal aims to produce more clarity in the reporting of reliability claims, as the information required by the checklist is necessary for interpretation of the findings.

Figure 3 Checklist for reporting reliability claims



Facets

Some studies, especially those using generalisability-theory, incorporate more than one facet of reliability. The facets incorporated and recognition of those excluded from the study should be indicated.

Conceptual

Comparison - Generalisation to other performances depends upon use of the term true scores in the study. As such, reports should be clear about the derivation of true scores and the extent to which the operationalisation of true score in the study is adequate. Use of true scores affects whether claims can be made about correctness of the scores.

Generalisation - To help with interpretation, reports should indicate what is known about the likely generalisation of the reliability estimate, for example, to different cohorts of test-takers.

This might include use of confidence intervals (Fan and Thompson, 2001) and previously published figures on the same or similar qualifications.

Design

Level - Where the study has been conducted at component level, the limitations of interpretation of the outcomes at qualification level should be documented.

Administration - Details regarding administration of the test, scoring and quality assurance should be given, and should be compared with the operational situation to establish the extent to which the study has ecological validity.

Representativeness - Interpretation of the reliability estimates is also dependent upon the extent to which the data were representative of the scores to which the reliability figure is to be generalised. As such, an indication should be given of the representativeness of the object of replication such as the raters. If the raters included in the study were not representative of the population raters, the reliability outcome would not be dependable. Similarly, the stimuli that has been rated needs to be representative, as does the data generated from those stimuli. Differing levels of variability in scores across populations (Lord, 1984) and proximity to cut-scores have an impact upon reliability estimates.

Method - The way in which the data have been analysed and the statistics used to report the reliability outcome should also be described. For precision, the equations for the analysis should be given.

Scale – The number of points, scores and/or grades used in the assessment has an impact upon the classification accuracy. Thus, these should be documented in any study.

Markers: Rater reliability

Consensus on the Reliability Programme emerged regarding reporting of correlations for inter-rater reliability studies; it was generally agreed that they are difficult to interpret unambiguously and therefore unhelpful as a sole reliability reporting mechanism. Classification consistency, classification accuracy, standard error of measurement or other forms of reporting should be used. Over time, it is possible that a consensus will emerge in the literature about which of these is most useful and dependable, but that is not the current state of the art and a variety of perspectives persist.

Consistency within an individual rater over different occasions (intra-rater reliability) is important, but has not formed the focus of research on this programme. Future research should investigate intra-rater reliability in English qualifications.

Research on the programme

Bramley and Dhawan (2010) presented data from the operational quality assurance of examiners from a range of GCSE (n=19), and A-level (n=24) question papers. The mean mark difference between Assistant Examiners and Team Leaders was less than one mark and close to zero for the majority of the question papers. Zero is the expected figure, as error is random; so positive and negative error values should be equally prevalent and cancel each other out to a mean of zero. Two question papers had mean differences of less than three marks. The standard deviation of these differences ranged from 0.19 to 7.16, with the largest value accompanying one of the two question papers with larger mean differences. These findings show consistent marking for most of the question papers; to the extent that those with more variability stand out.

Future research

Research should be conducted on inter-rater reliability on a wider range of examinations, so that work can begin on the setting of criteria for acceptability for inter-rater reliability statistics. Such criteria would assist examining bodies to identify assessments where the inter-rater reliability needed to be improved.

Regulation

Operational checks are conducted on the quality of inter-rater reliability for most assessments in England, and certainly for the large-scale public examinations. As such, large volumes of inter-rater reliability data are readily available for analysis. These data will not be in the same form across awarding bodies or examination types. However, Ofqual should require a minimum level of production of statistics on inter-rater reliability, such as the standard error of measurement. Where an organisation wishes to produce an alternative statistic in addition, they could do so and make the argument for it better representing inter-rater reliability for that assessment.

Recommendation 5 **At a minimum, the standard error of measurement should be produced to indicate inter-rater reliability for assessments regulated by Ofqual.**

Tests: Internal reliability

In general, reliability refers to the consistency of the outcomes of an assessment. As Hutchison and Benton (2009) point out, this is not about whether the correct things are measured or if the measurement is unbiased, but about the consistency regardless of the choices made in defining the framework for the assessment. This theory about reliability is embedded in a measurement theory that is often addressed as Classical Test Theory (CTT) but can perhaps be more accurately addressed as True Score Theory (Johnson and Johnson, 2009). In this theory a score of a student on a test is assumed to consist of a true score and an error term. The true score is the part of the test score that is based on the actual proficiency level of the student on the construct measured with the assessment, while the error term is a variation from the true proficiency level due to all kinds of error influencing the measurement. An additional assumption of CTT is that the error term is independent between administrations and between different assessments measuring the same construct. A formal description of how true score theory is related to reliability is given in Johnson and Johnson (2009).

The aim of estimation of reliability is to distinguish between the variance due to true scores and the variance due to error. Theoretically, this could be obtained by comparison of outcomes of repeated measurement or using a parallel measurement with the same properties as the original measurement. Practically, in educational assessment it is hard to satisfy the conditions that are necessary to get a good estimate from repeated measurements of the same test from the same candidates, marked by the same examiners; since it is often not reasonable to assume that students during repeated administrations of the assessment are not influenced by the first administration. Using statistically parallel assessments (see, for example, Gulliksen, 1950) to estimate the true score and error variance guarantees that the above issues will not occur. By definition the error terms of parallel assessments are uncorrelated and the measured construct is identical. But this only theoretically solves the problem. For situations where it is also not feasible to use a parallel version of the test to calculate the correlation between these test versions, procedures are available to estimate the reliability based on a single test administration.

Estimation based on a single test administration

Estimation of reliability from a single test administration involves dividing the tests into two or more parts and estimating the reliability based on the consistency between these parts. This is most clearly illustrated by the split-half method. In this procedure the test is divided in two halves that are assumed to be parallel. By calculating the correlation of these two halves the reliability of these test halves can be obtained. Using the Spearman-Brown “Prophecy” formula to predict the reliability of a lengthened test (Spearman, 1910; Brown, 1910) an estimate of the reliability of the total test is obtained. It should be noted that if the two halves in reality are not parallel the coefficient will underestimate the reliability. Furthermore it is obvious that in practice different ways to divide the test in halves will lead to different values of the coefficient. In the next section, we argue that it is useful to construct an equivalent halves reliability index, based on a split half where each half tries to match the other half in content and statistical properties. Both halves of the test can be seen as equivalent measures. The split half reliability of this construct can be interpreted as an equivalent form reliability, and in this way an estimate of equivalent forms reliability can be obtained from a single administration.

A variety of coefficients based on single administration is available, and all of them are lower bounds to the reliability; meaning that a low value of the coefficient does not necessarily mean that the reliability is (that) low, but if the value of the coefficient is high this is an indication that the reliability is at least as high as the coefficient.

Different single administration coefficients

A large number of different coefficients are given in both Sijtsma's (2009) critique on the use of Cronbach's alpha and in Hutchison and Benton (2009). The best known single administration coefficient is Cronbach's alpha (Cronbach, 1951). This coefficient is equal to the expected value of all possible split-half coefficients. Cronbach's alpha can be used with both dichotomous and partial credit items and is for dichotomous items identical to KR20 (Kuder and Richardson, 1937). Cronbach's alpha was not invented by Cronbach and was first described as λ3 by Guttman (1945) in comparison to five other single administration coefficients. Guttman's λ2 outperforms Cronbach's alpha in all circumstances, but is somewhat harder to calculate.

A coefficient that is at least as large as all other possible lower bounds is the so-called greatest lower bound (GLB) (Bentler and Woodward, 1980). The disadvantage of the GLB is that estimation needs to be done iteratively and estimates of the GLB can be positively biased for smaller data sets (Ten Berge and Sočan, 2004). Verhelst (1998) provides a procedure to correct for this bias, but more work on bias correction for the GLB needs to be done according to Sijtsma (2009). Alternatives to the GLB are given by Green and Yang (2009), Bentler (2009) and Revelle and Zinbarg (2009). Both Green and Yang (2009) and Bentler (2009) recommended the use of Structural Equation Modelling (SEM) to describe the behaviour on the test items and estimate a reliability coefficient based on the SEM.

Using a number of existing data sets, Revelle and Zinbarg (2009) showed that in some cases a coefficient called ω_t (McDonald, 1999) performed better. This coefficient is based on the factor structure of the items in the test. Revelle and Zinbarg also provide a reference to the *psych* package (Revelle, 2008) in the programming system R (R Development Core Team, 2008) which can be used to estimate ω_t and all of Guttman's coefficients: CRAN (the Comprehensive R Archive Network: http://www.R-project.org). Next to the above procedures IRT modelling can be used to estimate the error of measurement in a single test administration (e.g. see Hutchison and Benton, 2009). Based on the IRT model the information function and the standard error of estimation of the ability score can be determined, but it is also possible to construct reliability indices comparable to single administration coefficients (Linacre and Wright, 2001; Verhelst, Glas and Verstralen, 1993) and to estimate misclassifications based on IRT modelling (e.g. Hutchison and Schagen, 1994; Hutchison and Benton, 2009). Furthermore a sampling based approach can be applied in which the parameters from the IRT model are used to sample response patterns on a test. For example to estimate the standard error of measurement for a certain proficiency level on a specific set of items one could sample response patterns given the proficiency level, θ. The variance in the number of items correct between the response patterns can be used to determine the standard error of measurement. An overall statistic of reliability can be constructed as the ratio of the variance due to proficiency and the total variance that contains both error of measurement and variance due to proficiency.

A somewhat more complex sampling approach occurs if one also wants to take into account the variance in outcomes due to the variability in the content of the test. In this case sampling of a response pattern contains two phases. First, a sample test is generated using a sampling scheme from the pool of items that mimics the sampling process of the construction of tests. So

for example if it is reasonable to assume that the items are exchangeable from a content perspective, the variability of the content of the test can be modelled by sampling with replacement the necessary number of items in a test from the items available in the test. In the second phase a response pattern is generated on the sampled set of items (again using the item parameters of these items and the proficiency level θ). Clearly more advanced sampling procedures can also be used.

Internal consistency, generalisability and correlation between parallel assessments

Sijtsma (2009) argues that there is no clear relationship between Cronbach's alpha and the internal structure of a test. He shows that a 1-factor test can have any value of alpha, so alpha says very little about uni-dimensionality. Furthermore he shows that different tests with a varying factorial composition may have the same alpha value. From these observations it can be concluded that Cronbach's alpha should not be interpreted as a measure of internal consistency.

In the interpretation of reliability coefficients it is useful to distinguish between the reliability as the accuracy on a specific test or assessment in which the items are fixed or as the accuracy in which also the variability of the sampling of a set of items from the intended content domain is included. In the last situation the items are seen as random. Formally, true score theory has a focus on the first interpretation, while generalisability theory allows us to model the variance due to the sampling of items. Also structural equation modelling and IRT procedures can be used to estimate both coefficients that will or will not take into account variability due to sampling of items.

Research on the programme

The programme provides information about Cronbach's alpha, about classification error and about IRT-based estimation procedures, but no other reliability coefficients, like λ2, GLB or ω_t, are presented. Research into the parallel form reliability of the Key Stage 2 Science tests (Maughan, et al, 2009) also provides estimates of the internal reliability of the level 3-5 tests of 2004 to 2008. The reported values for the live tests were between 0.92 and 0.94 for total test scores. Cronbach's alpha coefficient for the components A and B and the anchor test ranged from 0.84 to 0.90. Similarly, He, Hayes and Wiliam (2011) reported values of Cronbach's alpha ranging from 0.91 to 0.97 for the Key Stage 2 tests in English, science and mathematics for the 2009 to 2010 test sessions.

Hutchison and Benton (2009) gave an extensive review of the different indices and procedures that can be used to report internal reliability. In the example data, they reported a Cronbach's alpha of 0.88 on the 2008 Key Stage 2 reading pre-test, while the correlation of the pre-test with the 2007 live test is 0.81, and with an anchor test the correlation was 0.85. While these results were somewhat lower than on Key Stage 2 Science test, these values are still considered a reasonably high reliability. They also provided the estimated accuracy of level decisions based on IRT modelling (83% correct) and the estimated consistency of level decisions (76%) between two parallel tests. The last value was substantially higher than the 70% found in the comparison between the live test 2007 and the 2008 pre-test. In comparison with the anchor test the consistency of the pre-test was 71% and in comparison with teacher assessment a consistency of 66% was found.

Wheadon and Stockford (2011) reported the percentage misclassification for a number of GCSE and A-level components. At least 89% of the students received a grade equal or adjacent to

their 'true' grade. Bramley and Dhawan (2011) provided a distribution of Cronbach's Alpha for GCE and GCSE tests. Approximately 25 to 50% of the test components had a Cronbach's Alpha above 0.85. The median of the reliability of the test components was 0.83 and the lower quartile of the distribution of the reliability indices was around 0.75. A small number (6) out of 287 components had a reliability index below 0.60.

Johnson and Johnson (2011) presented generalisability coefficients based on both relative and absolute measurement for a number of GCE and GCSE papers. The G coefficients based on relative measurement range from 0.87 for GCSE Drama to 0.50 for some GCE History optional combinations. GCSE Drama, GCSE Business Studies (Higher Tier), GCE Statistics A or AS had coefficients larger than 0.80. GCSE Biology and GCE General Studies A/AS and GCSE Music had coefficients ranging from 0.71 to 0.67, while GCSE French had a coefficient of 0.59 (raw marks) or 0.61 (adjusted marks). The coefficients for absolute measurement ranged from 0.44 to 0.84.

In summary, the research on the programme showed that the internal reliability of the tests and assessments was at a relatively high level. If the presented internal reliability indices were classified according to the guidelines provided by the *Commission of Test Affairs* (COTAN) of the *Dutch Association Psychologists* (NIP), the interpretation would be that the reliability of most NCT tests would be judged as good while only some of the tests would be classified as sufficient and no tests have a reliability that is found to be insufficient. If we evaluate the indices given in Johnson and Johnson (2011) and Bramley and Dahwan (2011) for the GCSE components, less favourable results are found. In Johnson and Johnson (2011), three out of eight GCE and GCSE tests had a coefficient above 0.80, two just above 0.70 and three below 0.70, while Bramley and Dahwan found some components with indices lower than 0.60.

In interpreting the reliability of the components, we must bear in mind that decision at the student level are not based on a single component but on the aggregate over multiple components. The presented internal reliability indices are often limited to Cronbach's alpha. Although this index is well known and often used, sufficient evidence is available to advise the use of different indices like the GLB, SEM and IRT based indices. Although the later indices are better than Cronbach's alpha, no undisputed advice can be given on the index to use. Depending on the situation, the order of performance of indices changes. From a practical perspective it is also relevant to consider that the SEM and IRT based indices involve a degree of knowledge about the underlying models of which it is unreasonable to expect that it is available in every assessment agency. Based on the above considerations, we advise the regulator not to specify the specific information or reliability index that needs to be reported. In this way the responsibility of providing adequate evidence of reliability remains with the assessment agency. This agency chooses the index and information that is made available and decides on the additional research that is carried out to provide information about reliability. A specification of an index could limit the quality of the provided information and would potentially discourage additional research into the reliability of the test and the assessment system.

Future research

As stated before the reported internal reliability indices in the programme were mostly limited to Cronbach's alpha, while the indices that were considered as better were not reported. Although the evidence about reliability based on Cronbach's alpha is positive it would be useful to have research available in which different single administration indices are compared. Most interesting is if practical situations and types of tests can be identified in which different indices

clearly perform better. Potentially this could improve reliability evidence and the procedure used in regulation.

A comparable research question is how the more complex and model based procedures like IRT, SEM and Generalisability Theory work when they are applied to research reliability for different assessments. A substantial amount of research on reliability based on generalisability theory was carried out in the programme (Johnson and Johnson, 2011). Also some research based on IRT was carried out (Bramley and Dhawan, 2010; Hutchison and Benton ,2009). But although Structural Equation Modelling in reliability analysis was recommended by Green and Yang (2009) and Bentler (2009), no structural equation modelling procedures were used to investigate reliability in the programme. In future research it would be useful to also evaluate the value of SEM for reliability analysis. Furthermore, a practical comparison between the different model-based procedures (IRT, SEM and Generalisability Theory) would give useful information to the test development agencies about the suitability of the different approaches to investigate reliability.

Recommendation 6 **At a minimum, a lower bound internal reliability index should be produced for each assessment. An equivalent-forms index would be preferable, where it is possible to produce it.**

Equivalent forms reliability

What assumptions underlie a judgement that two students with the same grade, but from different exam boards or from different sittings from the same board, are equally well qualified? This is the territory of *equivalent forms* reliability.

The idea behind this version of the reliability concept is that we expect the results that pupils get would be the same if they had taken a different form of the test. Indeed, this is assumed in many of the uses that are made of exam results; when a university selector considers a student with a B in this year's A Level to be better than one with a C in last year's, it is not just the equivalence of standards that is assumed but the equivalence of the exams too. More precisely the selector is assuming that:

a) the two tests measure the same 'thing', or trait, in such a way that candidates' scores would show the same rank order if they had taken both exams

b) grade boundaries are set so that the same number of candidates would fall in each grade.

Similar assumptions apply even within a single year, and a single examination. Thus, within a school, a teacher responsible for selecting or advising students starting an A Level course wants to trust that the pupils' GCSE grades would have been the same if they had taken a different version of the examination. The alternative, that getting into an A Level course is a matter of luck with the questions which happened to come up in this year's test, is not tenable.

The overall aim of exam and test developers is to maximise the validity of the results for the purposes for which they are intended to be used. Since most of these uses will involve decisions similar to these examples, equivalent forms reliability might be considered the most important form of reliability for educational assessment. It is unfortunate, therefore, that in our qualification system it is the least investigated.

Terminology – 'equivalent forms'

The term 'equivalent forms' is used here for the most general and most useful form of test equivalence, and is formally known as *congeneric forms*. The hierarchy of types of equivalence below are based on the descriptions given by Graham (2006),

> "**Parallel** tests are tests in which every test item is exactly like every other one, and there are the same number of items in each. Thus the expected score (called tau) for any student is the same on both tests, and so are the amounts of variability associated with them. In education, this is never a useful model for test equivalence.
>
> **Tau-equivalent** tests are slightly less strict. The item variability is allowed to be different for different items, but every pupil's score is still expected to be the same on both tests. Again, this is unreasonable for examinations.
>
> **Essentially tau-equivalent** tests are less strict again. Students' expected scores may vary, but the difference must be the same for every student; thus if a student is expected to score **X** on one test then they are expected to score **X**+**d** on the other, where **d** is the same for every student. The product-moment correlation between the two sets of scores is expected to be perfect, equal to 1. But this is still too restrictive for most educational tests.

Congeneric tests remove the requirement that the difference **d** is constant and, because the relationship between two sets of scores may not be perfectly linear, the product-moment correlation is not expected to be 1. Still, however, the expectation is that the rank-order of the two sets of students' scores will be perfect."

In this section, the phrase *equivalent forms* will be used to refer to *congeneric* tests or exams, since it is reasonable to expect that our examiners should aim to produce tests that would, in ideal circumstances, generate the same rank order of candidates. Perfect reliability of outcomes would also require the examiners to set the grade boundaries in such a way that the same numbers of students would get each grade, whatever form is used; the same numbers and the same rank order means, of course, that every candidate would get the same grade from every form. It is unreasonable to ask more than this from test constructors in a grade-based system. Since the setting of grade boundaries is not strictly a part of test reliability, equivalent forms of unreliability should be recognised as the most comprehensive measure of uncertainty in our educational tests. Specifically, the 'non-equivalence' of papers is the main source of what candidates mean by 'luck' in examining, and it gives the best overall estimate of the replicability of the scores on a test or examination component.

The various kinds of reliability

To understand what is meant by 'equivalent forms reliability' it is useful to adopt the notation of Generalisability Theory, or G theory (for a fuller treatment with more rigorous derivations, see Brennan, 1983; Shavelson and Webb, 1991). Instead of reliability, G theory focuses on the sources of unreliability, and uses the statistical methods of the analysis of variance.

1 *Unreliability due to random error*

Consider the simplest possible assessment, consisting of a single task where the responses are marked objectively (by a human or a machine). We can represent a person's test score in a systematic way as:

Eq 1a $$x_p = \bar{x} + \left(x_p - \bar{x}\right)$$

where x_p is the person's score, and $\bar{x}$ represents the average score of all the persons.

This equation is quite trivial, since the $\bar{x}$s cancel out, but its purpose is to separate out the term in brackets, which represents the *deviation* of this person's score from the average score of all the people. Squaring and adding up these deviations, across all the people, gives us the *variance* of the scores, and this is the basic quantity that is analysed in G theory. In this case the variance equation is trivial:

Eq 1b $$V_{total} = V_p$$

where the symbol V_{total} represent total variance, and V_p represents the variance amongst all the persons: $\bar{x}$ disappears because there is no variation in the overall average, since it's a constant. In this case there is no real analysis possible.

However, given the basic equation of true-score theory, that observed score equals true score plus error, $x = x_T + e$, we can make the underlying concept clearer by recasting the equation as:

Eq 1c $$x_p = \bar{x} + d_p + e_p$$

which shows the person's score as equal to the overall mean score, $\bar{x}$, plus a term specific to the person, d_p, which shows how far from the overall mean the person's true score is, plus a residual error term for that person. The error term is the only source of variability in this simple system, and the only source of unreliability. Now:

Eq 1d $$V_{total} = V_p + V_e$$

The total variance is partitioned into two parts, one the variance amongst the people, and the other the residual error variance. Now we can define the reliability coefficient as:

Eq 1e1 $$Rel = \frac{V_p}{V_p + V_e}$$

The numerator, the variation due to persons, is the 'true variance', while the denominator consists of the 'true' variance plus the relevant error term. For future reference this can also be written:

Eq 1e2 $$Rel = \frac{V_p}{V_p + \Sigma V_{error}}$$

with ΣVerror indicating that all of the factors that contribute to error are combined in the denominator, and that the more of them there are, the lower the measure of reliability is likely to be. In this simplest case of all, there is only one source of error.

Before continuing to more complex examples, it is worth emphasising that the word *error* here is a statistician's technical term. In everyday language it would be better called *variation arising from sources that we would prefer to exclude*. We want to measure reliably the differences between candidates, while any variation in their scores that arises from differences amongst markers or between test forms, or test conditions, or elsewhere, is undesired and confuses what we want to measure. *Error* does not mean *mistakes*.

2 *Unreliability due to inconsistent items*

The simplest reasonable test model will involve one other element, or *facet*. Suppose the test consists of several questions, but they are still marked objectively. Equation 2a shows the model:

Eq 2a $$x_p = \bar{x} + d_p + d_q + d_{pq} + e_{pq}$$

There are now five elements in the equation, representing the overall mean ($\bar{x}$), the person effect (d_p), and the residual error as before (e_{pq}), plus the effect of differences in the difficulty of the question (d_q), and an interaction effect from inconsistencies in how different persons perform on the different questions (d_{pq}). In variance terms this gives:

Eq 2b
$$\begin{aligned} V_{total} &= V_p + V_q + V_{pq} + V_e \\ &= V_p + V_q + V_{pq,e} \end{aligned}$$

The final two terms are combined in $V_{pq,e}$ since, in practice, there is no way to separate them in data. The reliability formula that results from this is the same as the alpha coefficient for measuring the internal consistency of a test.

3 *Unreliability due to markers*

Suppose there is a single task, as in part 1, but it is judged by markers, or raters (d_r). Equation 3a shows a model similar to the one in Equation 2a:

Eq 3a $$x_p = \bar{x} + d_p + d_r + d_{pr,e}$$

This looks like Equation 2a, but the interaction term this time ($V_{pr,e}$) refers to how different judges rate the same person's response. Unlike in the formula for internal consistency, this time the variance due to raters – V_r – *does* contribute to the error of measurement, since variation in the severity of markers is a key source of marker unreliability.

Of course, many studies of marker reliability will be more complex that this, since there will be more than one task, or question. In principle, then, the appropriate equation for the marker unreliability model will include both question and rater effects, as well as their interactions:

Eq 3b $$x_p = \bar{x} + d_p + d_q + d_r + d_{pq} + d_{pr} + d_{qr} + e_{pqr}$$

For a reliability coefficient, the numerator will still be V_p, but deciding which of these terms should be included as contributing to error of the congeneric kind needs care. In this case, since 'p' and 'r' both are, then so are all the interactions involving them, but the term that only involves 'q' is not, because we allow questions to vary in difficulty.[¶]

It's important to note that a reasonable model for marker unreliability will include error from the inconsistency of questions *as well* as error arising from the markers themselves.

4 Unreliability due to equivalent forms (simple)

In the simplest case, a model for 'equivalent forms' needs terms for variance due to persons (*Vp*), forms (*Vf*), and the interaction of these two (*Vpf*), and of course residual error (*Ve*):

Eq 4b $$x_p = \bar{x} + d_p + d_f + d_{pf,e}$$

With this model, the variance between forms is not a source of error (congeneric again) and the only source of error is the interaction term. This, however, fails to identify the factors that make up this error; in a typical case there will be variations between the questions within the forms, and variations between markers, as well as all the interactions between persons, questions, and markers. As with the marking case, we need a more complex model, such as:

Eq 4b $$x_p = \bar{x} + d_p + d_f + d_{q:f} + d_{r:f} + d_{pf} + d_{pq:f} + d_{pr:f} + d_{pfq:f} + d_{pfr:f} + d_{pfqr,e}$$

In Equation 4b the colons indicate that one factor is *nested* inside the other; thus *q:f* shows that the questions are nested inside the forms, so that a single question always appears in one form and never in any other one. The details of this model are not important, unless you are designing an experiment to explore it. What is important, is to see that the concept of equivalent forms reliability involves many sources of error, including contributions from both the internal inconsistency of the two (or more) test forms and the variability of markers, as well as the interaction of these two sources with each other, all in addition to variability caused by the forms themselves, and the interaction of items and markers with the forms.

The reliability coefficient, as before, will be of the form:

[¶] This model is intended to describe the natural context of exam marking. In a 'fully crossed' experiment designed to evaluate these various sources of error, some of these components may be 'designed out', leading to a slightly different formulation.

Eq 4c

$$Rel = \frac{V_p}{V_p + \Sigma V_{error}}$$

But this time most of the terms in Equation 4b2 will contribute to the error part in the denominator. From an analytic point of view this allows us to better understand what is causing the total amount of error we observe, and realise what we could do to reduce it. In equivalent forms reliability, the data comes from the administration and marking of two or more forms of a test, while estimates of internal consistency and marker reliability come from a single administration. There will therefore always be more 'error' in this type of study, and coefficients for equivalent forms reliability will always be the lowest of the three types (**Error! Reference source not found.**).

Table 3 Three levels of sources of unreliability

Type	Sources of unreliability
Internal consistency	questions, residual
Marking	markers, questions, residual
Equivalent forms	forms, markers, questions, residual

In this respect, of course, they come closest to properly recognising all the factors that reduce the reliability of educational assessment. In a practical exploration of equivalent form reliability, there may be a fifth significant source of variability that will further lower the apparent reliability. If the two forms are not administered simultaneously, each student may not perform equally well on the two occasions, some being better on the first, others on the second (Table 4). Therefore as well as the normal variance due to differences *between* persons there will be an additional variance *within* persons associated with occasion, and this will further reduce the observed coefficient of reliability.

Table 4 Equivalent forms sources of unreliability

Type		Sources of unreliability
Equivalent forms	(simultaneous)	forms, markers, questions, residual
Equivalent forms	(over occasions)	forms, markers, questions, occasions, residual

5 Further analysis of sources of variability

The main message is that there are many potential sources of unreliability that may need to be identified in any particular conceptualisation of reliability. It is important to distinguish two kinds of study. In an operational context we can only explore some of the sources of error, for example, since operational test data normally only covers the administration of one test per person it is not possible to estimate the contribution of the differences between forms to

unreliability. This does not mean that it does not exist and can be ignored, and it therefore also means that any reliability indices calculated from these data may be spuriously high.

In an experiment designed to investigate sources of variability, on the other hand, we can deliberately include more factors. The overall index will be lower, but we will be able to separate out the various contributions. This means that a well-designed experiment to study equivalent forms unreliability will also provide good evidence of the levels of marker unreliability and of test inconsistency. (For discussion of some of the possibilities, see Johnson and Johnson, 2009.)

We could include further sources of variability than consistency, marking and equivalent forms, and calculate a still lower estimate of reliability. Classification variables such as sex, school type, age or state of health could be included. In a quasi-experimental study, using the analysis of the variance, these could be used to tease apart even more aspects of what causes systematic variability in test scores. These kinds of analysis, however, probably go beyond what is normally considered to be reliability, and belong more to the realms of bias and validity.

On the other hand, unreliability of grading, where two equivalent forms may be given grade boundaries that are not *exactly* equivalent, will further reduce the reliability of the test outcomes, and will show up as reduced classification consistency. Since a grade boundary is normally an integer, and often the one nearest to the estimated 'true' equivalent score, grading unreliability is almost always a further source of unreliability.

Calculation of coefficients

As a precursor to this discussion it is worth considering what sort of coefficient is appropriate to measure what we mean by 'equivalent forms reliability'. The first point is that it is unreasonable to expect strictly parallel forms, or tau-equivalent, or essentially tau-equivalent forms, in most educational contexts. Congeneric forms are expected to maintain the same rank order of true scores, but without any common scale requirement beyond that; this seems a reasonable standard to measure our tests against because, in theory at least, we use boundary setting meetings to adjust for scale differences between different forms.

The appropriate type of correlation coefficient, then, is one that tests the stability of rank ordering, such as a Spearman rank order correlation. In theory, a product-moment correlation should under-estimate the reliability of equivalent forms, since it 'expects' one set of scores to be a simple linear transformation of the other: in practice, the effect may be too small to matter. Note too that a measure of classification consistency across two forms will also confound equivalent forms unreliability with unreliability in the setting of standards, which is an inevitable feature of a system which tries to equate two tests that are not perfectly parallel using integer scores and integer grade boundaries.

Adequate estimates of the *equivalent forms* reliability, free from other effects like *marking* and *occasions*, can generally only be derived from an experimental study. If two forms of a test generate sets of scores whose rank order correlation, or grade allocation, are not perfect the differences may arise in part from: internal inconsistency, marking variation, intra-person variation across the two occasions, inaccuracy in setting grade boundaries, or because the two forms were not congeneric. Coefficients obtained from non-experimental studies will not be able to separate these effects unless the required data happen to be collected.

1 ***Coefficient of equivalence***

There are more complexities, arising from the practical aspects of investigating the concept of equivalent forms reliability. The basic method is to set two forms of a test to the same set of pupils and to calculate the correlation between the two sets of scores; this correlation coefficient measures reliability as the *coefficient of equivalence*.

It is assumed, however, that there is no difference between the circumstances of sitting the first and second tests. Not only are the pupils the same persons; they are assumed to be unchanged between the two occasions. Usually this is achieved (it is supposed) by administering the two on the same day, one after the other. The order of presentation, however, may make a difference. Ideally, it should be randomised, meaning that a random half of the students will take the forms in the order A-B and the other half B-A. Except in an experiment, this will be difficult to do; suppose, for example, that one test is 'live' and the other merely serving as a pseudo-test to measure equivalence: can the order be randomised and still be fair to the pupils? And will the pupils be 'the same' on both if they know which the 'real' one is? Order may have significant effects, as pupils may be, for example, more tired when taking the second one.

2 Coefficient of equivalence and stability

If it is not feasible to administer two tests on the same day, they may be given on consecutive days, or with a gap of up to two weeks. The equation is then called a coefficient of equivalence and stability, and differs from the simple coefficient of equivalence in several ways.

First, a disadvantage with the simpler coefficient is that a pupil may be unusually poor or good on a particular day; if both tests are given the same day then the coefficient will be spuriously high. If school examination results are supposed to tell us how good pupils normally are at, for example mathematics, then we do not want to generalise from a single occasion, on which individual pupils may have performed unusually (even though that is what we do in fact do with our school examinations). In evaluating the validity of using these results for selection or accountability or any similar purpose we should prefer to know how stable the results are over a short period of time. In the days of intelligence and aptitude testing for 11+ selection, it was considered that a gap of two weeks ensured that any short term random variation in performance would be removed.

Not all variation over time is random, however. Pupils grow up, learn, revise and practice between the two occasions. If the gap is too long then these changes will spuriously reduce the coefficient, because there will be differences in the amount of change for different pupils. In effect, we are predicting the second test scores from the first set, and the longer the gap the lower the correlation. The two-week gap was judged to be the best compromise between these two threats that would spuriously raise or lower the coefficient.

The previous comments about 'sameness' apply again, to the coefficient of equivalence and stability. It is unlikely that two occasions a fortnight apart will both be thought 'authentic' by pupils in most circumstances.

Related to this is the validity issue of the 'shelf life' of examination results. Some tests, including what were formerly described as aptitude tests and some language proficiency tests, indicate that the test result should not be considered valid after a certain period of a few months or years. Estimating the stability of test scores over time, either by test-retest or equivalent forms, would help with setting an appropriate expectation of the shelf life of exams.

3 Coefficient of stability

This is not technically an equivalent forms reliability coefficient, but it is worth considering briefly the worth of estimating stability through a test-retest model, in which the same test is given twice, a week or two apart. Because the tasks or questions are the same 'questions', they are not a source of variability here, but there are circumstances in which this may not matter. Consider a competence test in a vocational setting. There may be a single criterion, for example to build a brick wall of sufficient quality in a given time, but knowing the task in advance will not unfairly benefit a candidate. The same may be true for a music performance. The concept of stability is still important.

Research on the programme

It is significant that there is no research at all on the programme into the reliability of equivalent forms in general qualifications or vocational certificate assessment. There are two papers that discuss and report the issue for national curriculum testing. Hutchison and Benton (2009) provided a description and explanation of the many different conceptualisations and measures of reliability and measurement error. To exemplify this they reported analyses of reliability for the 2008 Key Stage 2 reading test, based on its pre-testing in 2007, combined with data from an anchor test (used for standard setting), the live test the same pupils took in 2007, and teachers' estimates of the pupils' levels. The results are summarised in Table **5** (residual variance is omitted throughout).

Table 5 Reliability indices for Key Stage 2 reading (from Hutchison and Benton, 2009)

Coefficient		**Sources of unreliability**	**Value**
1	alpha	questions, markers	0.88
2	Equivalent forms (pre-test / anchor)	questions, markers, forms	0.85
3	Equivalence and stability (pre-test / live)	questions, markers, forms, occasions	0.81
4	Concurrent validity (pre-test / teachers)	questions, markers, assessments, occasions	0.77

Coefficient 2 concerns *equivalence* alone, while coefficient 3 concerns *equivalence and stability*. The anchor test is intended to be an equivalent form, and the figure of 0.85 can be taken as a reasonably accurate indication of the equivalence *in the low stakes context of a pre-test*. The figure of 0.81 in addition accounts for variability caused by both time and the change in conditions. The final figure, 0.77, compares scores on a 50 mark test to teachers' estimates on a 4 point scale, and must be considered more as evidence for concurrent validity than simply for equivalent forms reliability.

Maughan et al (2009) also presented data from pre-testing of Key Stage 2 tests, in this case five years worth of data from the science tests (Table **6**). A feature of these is that the test consisted of two 40-item sub-tests that were intended to be equivalent and both were administered, with an anchor test that was also equivalent to them, to groups of several hundred pupils a few weeks before they also sat their live test, which should again have been equivalent.

Table 6 Reliability indices for Key Stage 2 science (from Maughan et al, 2009)

Coefficient		Sources of unreliability	Mean value
1	alpha (80 items)	questions, markers	0.93
2	alpha (40 items)	questions, markers	0.87
3	Equivalent forms (form A,B / anchor)	questions, markers, forms	0.85
4	Equivalent forms (form A / form B)	questions, markers, forms	0.84
5	Equivalence and stability (A+B / live)	questions, markers, forms, occasions	0.85

It is remarkable here that the average coefficient of equivalence and stability was as high as the average coefficient of equivalence, despite the extra source of unreliability from testing on two occasions about five weeks apart and the change from low to high stakes between the two occasions. The individual values were, however, quite unstable ranging from 0.78 to 0.89 with no clear pattern between the different coefficients, and we should not make too much of this.

The overall average correlation for the various 40 item forms is 0.85, while their average alpha coefficient is 0.87. What is clear is that the correlations between the forms are almost as high as could be expected, given this level of internal consistency. Assuming no difference in grading standards, the choice of forms makes almost no difference to pupils.

There is no research of this kind so far for certificate examinations. In general there is no pre-testing of GCSE or GCE papers, and almost no pre-testing of questions for them. It is therefore difficult to think of an operational context in which equivalent forms reliability could be explored.

It is ironic that this is perhaps the most important form of reliability we can conceptualise, since it includes more of the sources of unreliability and more of the threats to validity than the others. Lack of equivalence between papers gives rise to what students so often see as 'luck', meaning that they did or did not get the questions they wanted. The implication is that, with a different form of the test, they would have done significantly better, or worse, *relative* to other students.

Future research

Are there any ways in which equivalent forms reliability could be more easily assessed? The problem is to find a way in which either or both of the two fundamental problems can be circumvented: a way of simplifying the data collection so as to reduce the cost, and/or a way of controlling the conditions so that the conclusions will be valid for live examinations. A few suggestions are made below.

At qualification level

The fundamental problem here is that, for the range of general qualifications, students are not allowed to enter for two examinations in the same syllabus specification at the same time. They are allowed to re-sit an exam a few months after the first sitting, but too much has happened between these two occasions to consider this as evidence of either equivalence or equivalence and stability. If anything, it would only show how much effect revision, re-teaching and coaching can have.

In the vocational field, there are some contexts in which immediate re-sitting is allowed (as, for example, with the National Certificate for Personal Licence Holders). In any case where this sort of 'double entering' is not forbidden it might be possible to arrange, with an appropriate experimental design, for a set of candidates to take two tests and be awarded a result based on the better of the two scores. Several test forms could be included, with each student taking two in a kind of matrix sampling design, to give a very good indication of the coefficient of equivalence under high stakes conditions for that qualification. An appropriate opportunity will come with the early administrations of the new functional skills tests.

At test level

As mentioned earlier, there is no pre-testing for most examination papers at present, but there are three possibilities that should be considered. First, there may be natural experiments that approximate to an equivalent forms experiment. In a sense this is the case when 'alternative written' papers exist for those who cannot or will not do practical papers, but these will probably provide rather low estimates, since the degree of real equivalence is rather poor. Again, optional sections or questions might seem to offer an opportunity, but it is likely that we would only find evidence of how similar the standards are in the options rather than how congeneric they are. Awarding bodies should be encouraged to look for contexts like the national curriculum science tests in which two components in an exam, preferably taken on different days not too far apart, could reasonably be considered equivalent in the congeneric sense. That is, where they test the same skills and do not draw on content that is likely to be taught or learned in different conditions, such as by different teachers or with different motivational effects.

Second, it is relatively easy in some circumstances to set up a 'natural experiment' that would allow equivalent forms evidence to be inferred. The equivalent forms coefficient is, like Cronbach's alpha, related to the very old idea of the split-half reliability coefficient, in which half of the questions are deemed to be one 'form' and the rest are deemed to be the other. Traditionally, a split-half coefficient was usually derived from the correlation between scores on the odd numbered items and on the even numbered ones, on the grounds that most of the possible spurious factors that might make two 'half tests' seem less equivalent (such as interest, tiredness, topic or difficulty) would be shared equally between the two halves; splitting the test in other ways would usually lead to halves that were less, rather than more, equivalent. The odd-even split-half coefficient would, therefore, be just about the highest possible correlation between halves in contrast to KR20 or alpha which is the average value from all possible splits.

Of course a rigorous odd-even split may not be the best way to generate equivalent halves, particularly in a paper where the questions have been designed to satisfy the requirements of a two- or multi-dimensional specification grid which classifies questions in terms of content, skills, and other properties. Awarding bodies should look for components within their qualifications where the questions can be split into two approximate halves in a principled way based on consideration of the content and skills involved in them. If this is to give a good estimate of equivalence, it is important that the splitting is done by a specialist, such as the Principle Examiner, to make the parts as equal as possible in every possible way. The correlation between scores on these two half-tests, augmented using the Spearman-Brown equation to adjust for test length (Spearman, 1910; Brown, 1910), would then give a good indication of the equivalence of two whole papers like the one analysed.

The weaknesses of the alpha coefficient were discussed in the previous section of this report, but if it is presented as an accompaniment to a split-half coefficient, as suggested here, the comparison is of particular interest. The 'equivalent halves' coefficient must be higher than alpha if the halves are in any way more than randomly equivalent. Since many of our examination components are designed to be composites of several distinguishable traits, it is likely that they will not generally be strictly uni-dimensional measures. Suppose there are two distinct traits in a test: most of the split-halves that contribute to alpha will be unbalanced with respect to them, and will therefore correlate less well than a split that is designed to be balanced in this way. In such cases the 'equivalent halves' correlation will give a substantially higher coefficient than alpha. Since the division into equivalent halves will reflect the regular process of generating a test from the syllabus specification, the 'equivalent halves' version of the split-half coefficient will also give a more accurate indication of the equivalent forms reliability of the real tests. Whilst we are aware that there are real practical difficulties with this approach for many qualifications (where there is little or no duplication of curriculum content), there will be assessment for which this is feasible. Additionally, experimental studies could be designed to investigate this, which would assist examining organisations in improving the quality of their assessment designs.

Finally, computerisation will lead to assessments that use randomly parallel sets of items or tasks, and where the marking – if any – is spread across many markers for each candidate. In this case, every candidate will take a different form, the coefficient of equivalence will be the *only* form of reliability that matters, and it will be imperative for it to be estimated. It is common for this to be done for computer adaptive tests, for example, by allowing candidates to take the test twice; the first form may be considered as a 'practice' to make candidates familiar with the format of the test, but it still gives a reasonable lower bound for equivalent forms reliability. New formats of assessment are likely to offer still more opportunities and demands for assessing the equivalence of purportedly congeneric test forms. It may be that new ways of conceptualising equivalence, and reliability, will be needed to evaluate them properly.

Regulation

Reliability measures can be placed in a hierarchy of five levels, according to the sources of unreliability that they include in their estimates (Table **7**). At Level 5, reliability is an estimate of the replicability of the grades or other outcomes that students are awarded, and as such is appropriate to indicate the overall reliability of an assessment system, such as 'A Level History'.

Table 7 Hierarchy of reliability indices

Level	Reliability measure	Sources of unreliability
5	Outcome Consistency	questions, markers, forms, occasions, grading
4	Equivalence & Stability	questions, markers, forms, occasions
3	Equivalence	questions, markers, forms
2	Marking reliability	questions, markers
1	Internal consistency	questions

Level 4 differs in that it does not include differences in the standards applied when test scores are converted into grades. It is the appropriate measure when the question is how reliably a specific examination or test rank orders a group of students. The coefficient of equivalence and stability takes into account that each student may perform differently on different tests and on different days.

Level 3 is very similar, but it addresses the reliability with which an examination or test rank orders a group of students on a particular day; it still takes into account that they might perform differently on a different, but supposedly equivalent, form of the test. Since our educational system assumes that candidates will know exactly when the examination will take place but will not know exactly what questions they will be asked, this is the level of reliability measure that seems to fit most closely with the assumptions of the system.

From a validity perspective, Level 5 probably fits most closely with how assessment is *used*, when exam results are used to select the 'best' students for university, sixth form study, or employment, all five of these sources of unreliability are relevant. But when the purpose is *quality assurance*, things may be different. Level 1 is a minimal level; failure to achieve adequate internal consistency in an assessment component means that, in effect, there is no proof that anything is being measured well. A failure to achieve adequate marking reliability at Level 2 means that there is something measurable, but the procedures being used fail to measure it well (so long as we have good evidence that level 1 was passed). If both of these Levels are adequately achieved, students have been measured successfully on a single occasion with a single form of the test. Level 3, the coefficient of equivalence, checks further that the measurements obtained can be generalised to different versions of the test, and need not be re-confirmed for every single form that is used.

To go further, to Level 4, is probably going too far. As discussed in the introduction to this report, some occasion-related factors are dealt with through special consideration procedures, whilst others are considered 'the luck of the draw' in our educational culture. Important uses of qualification results, such as selection for Higher Education, require that they tell us about a stable characteristic of students. However, we know that students will perform more or less well due to occasion-related factors such as a headache that will not be given special consideration. Our assessment systems would have to be re-designed if we really wanted to deal with this adequately and assessments would have to be made of each individual over repeated occasions. But then we run into the problem that individuals change over time – they forget and they learn. So the time difference between occasions matters, and larger gaps would suggest that our assessments were less reliable than if we investigated our assessments within a short time period (even though this effect was more to do with changes in individuals than the reliability of our assessments). Some sources of occasion-related error are therefore tolerated by stakeholders in the English educational assessment culture (Ipsos Mori, 2009, p.19; Chamberlain, 2010, section 5.3; He, Opposs and Boyle, 2010, Figure 11).

Recommendation 7 **Ofqual should gather evidence of equivalent forms reliability for a range of qualifications, since this is the most comprehensive measure of assessment reliability. This may require a designed experiment and the findings will indicate whether the three sources of unreliability included in a coefficient of equivalence are large enough to invalidate the likely uses of the test.**

Recommendation 8 **As part of Ofqual's qualification accreditation process, awarding organisations should be required to demonstrate adequate levels of equivalent forms reliability. Sources of unwanted variation could result from aspects of the design that are not controlled by the awarding organisation or Ofqual, but**

technical information can then inform the discussion with other parts of the system, such as the Department for Education.

Teacher Assessment

The programme has not sponsored studies of the reliability of summative assessment by teachers to match, in technical detail and quality, its other studies of external examinations. Given the many different aspects of reliability competing for attention, it was judged that this one should be deferred. One justification for this is that most of the issues to be tackled for conventional tests are also present for this area, but are compounded with a range of problematic issues which are peculiar to assessments by pupils' own teachers and schools.

Here, we elucidate a basis for considering future work rather than reporting on progress already achieved. Given the difficulties, it is necessary to first consider whether this area merits investment in the specific types of enquiry that it would require. This issue will be discussed in the next section, followed by an account of the history of change in the systems that have set the framework within which assessments by teachers have contributed to public examination results. Next, we consider the available evidence about the potential reliability of assessments by teachers, with particular attention to evidence about the teacher assessed contributions in public examinations in the UK. These lead to a discussion of the various threats to reliability that arise in the context of assessments by teachers and schools. Finally, we set out some suggestions for future work.

Importance

A recent report drew attention to the expansion of teachers' roles as standards-based systems are being implemented in many countries, commenting that

> […] the teacher is increasingly being seen as the primary assessor in the most important aspects of assessment. The broadening of assessment is based on a view that there are aspects of learning that are important but cannot be adequately assessed by formal external tests. These aspects require human judgment to integrate the many elements of performance behaviours that are required in dealing with authentic assessment tasks.
>
> (Stanley et al., 2009, p.31)

This statement reflects a longstanding concern that assessment should change to reflect broader aims of education. Recent documentations and debates in the E.U. Council of Education Ministers have identified the importance of developing 'key competences', and are exploring the problems of reflecting these in national assessment systems:

> Key competences are a complex construct to assess: they combine knowledge, skill and attitudes and are underpinned by creativity, problem solving, risk assessment and decision-taking. These dimensions are difficult to capture and yet it is crucial that they are all learned equally. Moreover, in order to respond effectively to the challenges of the modern world, people almost need to deploy key competences in combination.
>
> (E.U. Council of Ministers, 2010, p.35 section 6)

These statements focus on the validity of assessments. Reliability is a necessary but not sufficient condition for validity (Crooks et al., 1996), and optimisation then requires a trade-off between the two. Ofqual's policy on validity has only been expressed hitherto in very general

terms and further detailed development will be needed. (The general policy can be found at http://comment.ofqual.gov.uk/regulatory-framework-for-national-assessments/).

It is clear that some important aims in the curriculum cannot be assessed within the limits of tests which are externally set and marked and taken within the limits of formal test occasions. Whether these aims are so important that a national system should include the methods of assessment, with their special administrative and other resources, to reflect them is a matter for debate. Another recent report has added a further consideration by arguing that

> […] teachers can sample the range of a pupil's work more fully than can any assessment instruments devised by an agency external to the school. This enhances both reliability (because it provides more evidence than is available through externally devised assessment instruments) and validity (it provides a wider range of evidence).
>
> (Mansell et al., 2009, p.12)

The results of MacCann and Stanley (2010) discussed below provide some evidence about this claim. From a broader perspective, arguments can be made for giving increased emphasis on supporting and developing the status and quality of teachers' summative assessments:

- the expertise teachers may develop by participation in national/public assessments will have more extensive benefits, because for most of the years of schooling, teachers and schools have to conduct their own summative assessments, the results of which can have important effects on the progress of their pupils:
- the difficulty teachers have in aligning good formative interactions with their pedagogic practices as whole is exacerbated if they play no part in the summative assessment system by which they and their pupils are judged.

Due to the systemic issues outlined previously in this report, it is difficult to specify precisely the limits to Ofqual's remit in teacher assessment.

Use of teacher assessment

Two broad areas are relevant to the present task, the first is to do with the assessments teachers and outside agencies make during several stages of schooling, the second is concerned with the terminal public examinations which are high-stakes for all pupils. For the former, there are teacher assessment programmes organised, in England, by the Department for Education (but dependent on the future of the work of the National Strategies). For the latter responsibility lies with the awarding bodies working within national guidelines laid down by the QCDA.

In the first of these programmes, the assessment policy developed following the 1988 Education Reform Act required that teachers' assessments form part of reporting achievements at ages 7, 11, and 14, but laid down that the results on national tests should be reported independently. This decision undermined the status of teachers' assessments; when, at a later stage, teachers were allowed to submit their assessments *after* they had been told the results of the external tests, comparisons of the independent alignment of the two sets of results was compromised. The more recent replacement of the tests at age 7 by scheme of teacher assessment has been evaluated for its feasibility and acceptability, but not for its reliability, whilst a new scheme for the age 14 assessments is still under consideration.

It should be noted here that the phrase 'teachers' assessments may refer *either* to their judgment on a specified task or tasks implemented within the normal classroom teaching, *or* to judgments based on overall impressions gained from class-work, homework, participation in class or group discussions, and so on. In the latter case, there is no objective evidence that can be exchanged to support the judgements. Researched data on reliability is largely from the first of these - for obvious reasons.

However, it is significant that, for England, the National Strategies unit has developed the Assessing Pupils' Progress (APP) plan. This involves the production and trials of resources and training support to develop frequent assessment by teachers between years 1 and 9. The outcomes were intended only for internal use in schools, but it was emphasised that the approach should be as a whole school strategy. Sets of exemplary test times were also provided, developed as 'performance items' taking advantage of the absence of formal testing constraints. Whilst many have found the system, with the support provided, to be very helpful, others have raised two problems. One is that the emphasis on judging each of a large number of detailed performance targets can lead to an atomised approach in teaching (whereas real-world tasks require judicious selection from a set of available items of knowledge and skill). The other is that there is disagreement about claims by some that this system integrates with assessment for learning: claims that it is more than a system of frequent testing. For example, a statement agreed by 31 delegates from 6 countries at the Third International Conference on Assessment for Learning (Klenowski, 2009) states that assessment for learning 'has been (mis)interpreted as an exhortation to teachers to (summatively) test their students frequently to assess the levels they attain on prescribed national/state scales in order to fix their failings and target the next level. In this scenario, scores, which are intended to be indicators of, or proxies for, learning, become the goals themselves. Real and sustained learning is sacrificed to performance on a test'. The Director of the National Strategies wrote that, 'I have tried very hard to leave readers in no doubt that APP exists to help with summative teacher assessment. It is not AfL. The supplement does provide an argument that an improved knowledge of progression can help teachers make better choices about next steps in learning.' (*personal communication*, July 2010).

In a detailed account of the APP programme (Stanley et al., 2009), attention is drawn to the similarity of its structure of guidance and of its support materials for teachers?, to the systems in three states in Australia. However, there is a significant difference, for in the latter each system is part of an integrated state policy which includes the terminal examinations for the public certification of students, whereas in England there is no link because the Awarding Bodies are involved only in the terminal examinations.

The second part of this section will focus mainly on GCSE and GCE assessments. This discussion will only consider work in England, although it will thereby explore principles of more general application. It will also omit consideration of vocational assessments with their workplace emphasis. The report by Johnson (2010) on which the account here draws, includes brief accounts of systems in Wales, Northern Ireland, Scotland, and of vocational assessments. All three of these countries have relied more heavily on teacher assessment, and have abandoned, or never had, formal blanket testing pre-16, with a change in Scotland in 2005 to a sample-based monitoring programme (the Scottish Survey of Achievement).

In the 1960s and 70s, the high-status School Certificate 16+ examination included a 100% teacher assessed option in English (Smith, 1978) and teacher assessment played a large part in the several of the local developments of the lower-status Certificate of Secondary Education.

When these were replaced by the single GCSE (General Certificate of Secondary Education) system, coursework was included as a significant part of the new system, with emphasis on assessing 'skills not easily tested in timed, written examinations' (QCA, 2006). However, no attention appears to have been paid to its reliability. These public examinations had a chequered history (Whetton, 2009). Coursework components had, at various times, different weights between different subjects, with one 100% option in English. Whilst many have claimed that teacher assessments should be included in high-stakes assessments, some teachers have disagreed.

Whilst teachers of English have regularly supported their involvement through coursework assessment, mathematics teachers voted in 2005 for abandonment of the coursework component (QCA, 2006). In a more recent exploration, some mathematics teachers explained that they accepted that realistic maths problems could not be assessed in written tests (as argued in ACME, 2005) but that the constraints of the moderation system had driven them to make such work into safe, and so stereotyped, exercises (Black et al., 2010). A similar finding emerged from a Royal Society enquiry with science teachers who felt that they had to use stereotyped practical investigations, repeated year upon year, to ensure that good results emerged (Black et al., 2004), so that the work lost its intended validity. Thus, it is not easy to compose a system in which the potential for enhanced validity will not be undermined by accountability pressures.

General concern about coursework led to a comprehensive revision in 2006. The basis for the present practice at GCSE (QCA, 2007, 2009) classifies subjects in respect of their coursework component into 60%, 20% and 0% categories.[5] All coursework components have to fit in with a scheme of 'controlled assessments'. The scheme is based on three stages of the conduct of any coursework: task setting, task taking, and task marking, the latter involving both the school's teachers and external moderation. For each of the first two, Awarding Bodies have to specify one of three levels of control: with a high level of control, for example, of task setting, the school would have little freedom in the choice of tasks, but this may be offset by very flexible rules about task taking. For the actual work on the task, the numbers of school class hours used, the possibility of group collaboration, and the actual individual production of the product (as opposed to preparatory research) might all be controlled at one of these levels, whilst for the actual marking, the control would have to be medium or high. The reports by the Joint Council for Qualifications (JCQ) in 2010, and by Johnson (2010) provide more details. Since this new system will be the norm for all GCSEs in England, Wales and Northern Ireland, it sets, for the present, the context for future studies of reliability. This context, and the reliability challenges that it presents is discussed in more detail later.

What does published evidence tell us about the reliability of teacher assessment?

Evidence from UK practices

The research evidence about teachers' assessments at the end of Key Stages 1, 2 and 3 will not be reviewed here in any detail. Impending changes mean that the relevance for any future

[5] Examples of subjects with 60% for controlled assessments are art, modern foreign languages and music; with 25% are English literature and all sciences and humanities; with external testing only are mathematics, psychology and law.

system is hard to judge. It should be noted however that studies relating such assessments to the Key Stage 2 test results have shown that the two sets of results agree to within plus or minus one level for over 99% of the pupils samples investigated, with about 75% of the levels judgments being the same (Tymms, 1996; Reeves et al., 2001). However, the number of levels involved at this stage is small (effectively levels 3, 4 and 5) - it would be more challenging to achieve such agreement at the end of Key Stage 4, for example.

A study by Taylor (1992) of coursework assessment in AQA entries compared the assessments of samples chosen from three GCSE subjects (English, history and mathematics) and one A level subject (psychology). For each subject, judgments by three moderators (one being the original moderator) were compared with one another, and with the original assessments returned from the candidates' centres. The very detailed analysis of the data showed that correlations between the moderators and between their judgments and those of the centres lay in the ranges 0.87 to 0.97 in English, 0.75 to 0.94 in history, 0.91 to 0.97 in mathematics, and 0.73 to 0.95 psychology. Most of the comparisons (nine in each of the four subjects) showed that the differences were statistically insignificant. Overall differences of judgement about the rank-ordering of candidates between centres and their moderators were not markedly greater than differences between the moderators themselves, a finding which calls in question too strong a reliance on the judgment of a single external moderator.

The moderators marking would have led to some changes in coursework grades, with between 14% and 20% lowered and between 4% and 9% raised, other studies also report such over-marking by teachers. It should be noted that these results were in a context in which the choice of coursework task was left to the centre, and the mark schemes was set out in broad guidelines. In general, the author judged that the numbers of discrepancies found were comparable with the results available at that time from investigations of the multiple marking of written paper scripts.

A questionnaire completed by those involved raised several general points about the system, which are relevant to the arguments about quality discussed below. Notably,

- Internal standardisation in centres was satisfactory, except in FE colleges where the work was often in the hands of part-time/evening staff dealing with large classes.
- Most centres set appropriate tasks, but in history and in mathematics some tasks did not give candidates opportunity to demonstrate higher-level skills.
- Centres valued the advice given by moderators as part of their task, and where their feedback pointed out unsatisfactory features this led to improvement in the next round of moderation; moderators emphasised the importance of their feedback responsibilities.
- Moderators could not take account of the conditions in which work had been produced - the product was the sole basis of their judgment; likewise, it was hard for moderators to take account of such 'ephemeral' skills as ability to discuss ideas orally.
- A few moderators were concerned that selection of their moderation sample by centres might be open to abuse.

Other types of study have explored the alignment between teachers' assessments and the results of external tests. One example is the analysis by Johnson and Munro (2008) comparing such results in the Scottish system: teachers assessments were more generous, but they also differed in that their score distributions were more bunched than the external test scores around the 'expected' levels of the national curriculum. Moreover, these effects varied

significantly between the school subjects. Similar results for the Key Stage tests are mentioned above.

Evidence from other countries

A review by Harlen (2005) identified twelve research studies of the reliability of teacher assessments, only seven of which were judged to be of high quality. Given that these were varied, in the pupils' ages, in the school subjects explored, and between inter-rater agreement and agreement of teachers' ratings with external test score results, it is not possible to arrive at any general conclusions, particularly as some showed good agreements and some did not. This varied pattern of the results was confirmed in a more recent review by Stanley et al. (2009). The main value of these reviews was that they help to identify some of the criteria which have to be met to achieve high reliability.

The sets of five tasks for English and psychology reported in Taylor's 1992 study make these cases examples of the portfolio approach[6]. A comprehensive survey of trials of the development of portfolio-based assessments, in grades 4 and 8, for writing and mathematics in two US state and one US city systems (Koretz, 1998) showed that inter-rater correlations were far too low in the first years, but rose with experience and training, to values of 0.8 to 0.9 in mathematics, and 0.6 to 0.7 in writing. The inter-correlations between the different tasks were low and two generalisability analyses showed, in one state, that the largest contribution to the variance in mathematics (about 25%) came from student-task interaction, whereas in writing this interaction accounted for only 7%. It is hard to compare the different results, partly because differences in the specifications of the portfolio contents, and partly because the rules for selecting only one's 'best piece' for assessment, and the rules for separate assessment and/or combination of different components, varied between the different systems. For validity, correlations with related components of test scores were very low, although how high such correlations should be is a matter for debate: the use of non-standardized tasks, weak guidelines for the inclusion of evidence in portfolios, and inadequate training in marking, were all identified as causes of difficulty. Similar findings were reported in a smaller-scale study of another US region by Shapley and Bush (1999).

Better results were reported in the Stanley et al. (2009) review. Of particular interest are the results from two of the Australian states. Good inter-rater agreement was reported from a study in Victoria, and for portfolio assessments, Masters and McBryde (1994) reported high consistency in double-marking studies, with inter-correlations of 0.94 and very few results disagreeing by more than one achievement level. In both of these states, there have been state programmes, sustained over many years, to give structure and support to teachers' own assessments.

Another study (MacCann and Stanley, 2010) used data from the school-leavers examination in New South Wales to explore classification consistency. In this examination, teachers' assessments and external test results contribute equally to the final grade. The study concluded that the moderated teachers' assessments classified more accurately than the test results; for example, in English 85.9% were classified accurately by teachers, 81.9% by the tests. The corresponding results in mathematics were 94.9% compared with 92.4%. The authors argue

[6] For the controlled assessments scheme, Awarding Bodies have freedom to specify the numbers of tasks, and it is not yet clear what these numbers will be, given possible diversity between the Awarding Bodies and between different subjects.

that it is the detailed care with the teachers' assessment scheme that produces these results, but question whether this modest gain justifies the extra work that the system might be expecting from teachers. This point is only cogent if such work is justified merely as improvement in reliability

In these Australian examples, it is clear that any procedures for the alignment of school-based assessment results with those of external tests have important effects when such assessments count for a percentage of the overall high-stakes results of individuals. A linear scaling may be used to adjust for differences between, for example, schools in overall mean values (given good agreement between the rank ordering). Non-linear relations between two sets of scores were used in some cases: if one set is regarded as the bench-mark, then more complex scaling may be used to secure alignment of the other set.

All of the above evidence relates to reliability. Research studies of the validity of teachers' assessments present more complex problems. Since the purpose of these assessments is to measure qualities not covered by formal tests, evidence of alignment with the results of such tests cannot be unambiguous evidence of validity. Audits must rely either on predictive validity or on analyses of construct validity. Judgments that compare task demands with curriculum aims may be part of the latter, but where such aims as real-world simulation, or some of the aims described under the EU's key competences are at issue, ways to operationalise the construct definitions in concrete tasks will be problematic.

Nature of the problem

The term 'coursework' covers a wide range of possible activities. These can range from informal notes about each pupil's achievements made by a teacher during classroom work primarily conducted as a learning exercise, through to a written test taken under conventional test conditions to check progress for example, at the end of a topic or module. It can also include a single piece of work assessed by the classroom teacher, or a portfolio of several pieces of work which might be assessed by a school scheme which includes moderated group assessment by several teachers. The range of variety is such that it is not meaningful to make general statements about the reliability of teacher assessed coursework. It also means that is hard to generalise from published research, even if it were extensive (which is not the case), for each study is inevitably about one specific system of assessments by teachers. So it is necessary to set up a taxonomy of the various features or stages which might help categorise and compare different schemes and systems. The following draws on the 'controlled conditions' specifications, but adds extra detail based on the specifications set out by Harlen (2005), in Johnson's (2010) report and in exploratory work with teachers by Black et al. (2010). The main outline is simple: a task is designed, it is presented to pupils, they engage in the task, their work leads to a 'product', this is assessed, and the result reported. However, the variety of possibilities is linked with a variety of threats to reliability, and calls therefore for a more complex account, as follows:

• Design of the tasks – issues here are:

- control – who specifies the tasks;
- validity – which curriculum aim or aims is each task meant to assess;
- discrimination – do tasks allow some opportunity for the lowest achievers to show what little they can do, yet offer challenges to discriminate between the high achievers;
- repetition - whether or not tasks should change from year to year;

- comparability – do different teachers use (some) common tasks to help in moderation.

• Presentation to pupils: this could be rigidly controlled, e.g. using a text or a set oral form, or left to each teacher's judgments so that presentation can be adapted to the attainments of the class, and can be adapted if a teacher judges that particular groups or individuals may not do themselves justice because they do not understand what they are being asked to do.

• Carrying out the task – here there are many possible variations:

- group work, or individual work, or a sequence e.g. with group research and sharing of resources, followed by individual production;
- access to resources, written, other media, friends and family, and whether during preparation only, or in production also;
- context – in set class times, or at and between such times;
- separation into preparatory and production phases, with control over conditions possibly loose in the preparation phase, and tight in the production phase;
- timing – both in the length of time allowed and the stage in a module or course when it is produced

• The product which is to be assessed: this may be ephemeral (e.g. a drama presentation), or an artefact (e.g. in art or technology) or a written account. It may be a single piece, or a portfolio of several pieces of work with tight or loose rules about the variety envisaged (e.g. realistic problems in maths varying in complexity and in the of skills required).

• Assessment of the product: obviously, this has to be matched to the nature of the product, but relevant features will be:

- the criteria or schemes that are used, which ought to match those of the curriculum; the level of detail in such schemes, and the balance between analytic and holistic approaches matched to the differing nature of different subjects; the level of detail at which criteria are specified; rules for the aggregation of component marks - with hurdles for some components where appropriate.
- those undertaking the assessment: the pupil's teacher, another teacher, several teachers;
- checking and moderation processes intra school, and/or inter school, and externally by the AB; how samples for moderation are selected; whether or not scrutiny or moderation is by blind marking or with marked products; what procedures are used to resolve or follow-up discrepancies.

• Procedures to reduce the effects of any teacher's personal bias, which may involve ad hoc training and a clearly specified moderation process.

• Strategies to limit and detect plagiarism, and copying between pupils.

This seems a formidable list. For the 'coursework' assessment in any one subject, a particular Awarding Body will specify a set of rules which will limit the possibilities. The guidelines in the new controlled assessments scheme, although narrower in scope than for previous coursework schemes, might mean that the sets of levels of control may differ between Awarding Bodies in the same subject. The modular system also introduces extra variability in the times and stages in a course when the assessments are produced, and therefore in the tasks used. As outlined in the introduction to this report, tight control over all features can be counterproductive, as it can

undermine those elements of validity which were the reasons for undertaking this type of assessment.

Two features of the above list deserve special attention. Any moderation process in which teachers are involved should first require that teachers within a school, or a consortium of schools, check their judgments by circulating sample scripts for marking, preferably by blind marking, and thereby both arrive at consensus results which will reduce some of the threats to reliability, and also develop a shared understanding of the criteria. In so doing, it is clear that it is the school, rather than any individual teacher, that should take responsibility for 'teachers' ' assessments. A QCA report has commented that

> Standardisation within a centre is required and there is much good and often very thorough practice taking place. However, internal standardisation is not apparent or consistent across all centres. Awarding bodies need to carry out further checks and provide better guidance.
>
> (QCA 2006, p. 11)

The short training meetings which Awarding Bodies generally provided, involving between one-half and one whole day, have to cover the administrative and general briefing involved. There may be little time for practical exercises of the type which are regarded as essential in training examiners of external tests. Such work within schools takes time and raises work-load problems. Teachers will be reluctant to take such work seriously unless the work is of value for their teaching. An exploratory study by Black et al. (2010a) aiming to improve the internal summative assessments in three schools, exposed a need for teachers to develop expertise.

However, collaboration between teachers should address all aspects of a school's policy for summative assessment. That is, it should involve agreement on such matters as uniformity or diversity in choice of tasks, in timing, presentation, and so on, as well as in moderation procedures. It is a common experience that teachers have found such investment of value for their teaching and as a resource for their professional development in general. The accounts of three controlled assessment tasks given by Johnson (2010) illustrates both the varied nature of these, and the way that each shows the selection and co-ordination of a range of concepts and skills. These are clearly valuable learning opportunities and not merely assessment tasks which interrupt a learning programme.

The policy of the state of Queensland is a model in this respect: the certification for school-leavers at subject level is based entirely on school-based assessments. Intra- and inter-school collaboration starts with consensus to approve each schools work programme at the start of a year and progress on this programme then has to be monitored half-way through a course. When portfolios are produced, their marking has to be checked by group consensus, and then samples are externally reviewed. Thus strong support for teacher assessment is the basis of the state system: it is not an extra to an external testing programme. Three other features common to this and to other developed systems are: emphasis on standards based assessment, with curricula which set out criteria in a scheme of progression; an expanded role for teachers which gives them more responsibility for the task; more roles in the setting of standards and significant responsibility in moderation; and an emphasis on expansion of support for teacher assessment through model assessment tasks to help teachers internalise assessment standards, supported also by task banks and strong training programmes (Stanley et al. 2009).

Ways forward

Many of the problems of principle, in the structure and systems, and in the national policies that set the context for the development of school based assessments lie outside Ofqual's remit. However, its technical advice can help inform debates and decisions. One conceptual problem is that the classical definition of reliability is focussed on the notion of reproducibility. It is not about fairness in general. In respect of the many variables which affect individual performance, some pupils are bound to be advantaged in comparison with others whatever the mode of assessment. Such effects may be more significant in school-based assessments than in formal tests. For example, the choices in timing and in levels of control for controlled assessments all create variability which might give advantage to some, for example of having a more effective teacher or a clearer school policy, and thereby make lack of fairness more directly evident. The rules and procedures for public assessments have to be framed to minimise such effects, but if they are made so tight that teachers are effectively merely supervising and marking an externally composed test, or a specified set of test items, the exercise may have little advantage over formal testing. So one technical problem is to find ways to explore the trade-offs between validity, reliability and equity which different levels of control may involve.

A further issue is raised by the claim that assessment by teachers and schools can be more reliable because they have numerous opportunities to observe pupils and to examine the work that they produce. A system may use this advantage in two ways: one is to focus on only one or two main products, but to allow latitude in judgment of these so that the teacher can use the fact that she/he 'knows her/his pupil'; this is almost certainly unacceptable. The alternative is to have judgments based on a collection of pieces of work produced over many occasions and involving a variety of types of activity. This would raise quite specific problems, about the degree of uniformity for the contents of portfolios, of pupil's involvement is selecting or assembling their own portfolios, and of aggregation and grading.

It would clearly be helpful to encourage a stronger set of research exercises in this area. With the swift pace of change of regulations in this area, conducting research that will have relevance in the future is problematical. The current GCSE may be an obvious context for such research as the controlled assessments system settles down. However, the number of variables is daunting: the different school subjects, the different choices between levels of control in the various stages (task selection and production, marking, and moderation), and the variations in timing between GCSE modules, would all have to be considered. Sample groups would have to be selected for a degree of uniformity in at least some of these aspects: information, perhaps from the Awarding Bodies, about how schools were making some of the choices, would help to guide such selection.

Given that a great deal of operational data is collected on many public examinations, we make the following recommendation.

Recommendation 9 **Statistics on the reliability of teacher assessment should be produced by awarding bodies.**

Workplace Assessment

Oftentimes during the work of this programme, researchers working in the field of work-place, or more broadly, vocational assessment, would respond to the discussions by claiming that what was being said about reliability did not apply in this context. The view of the Technical Advisory Group is that the issues in work-place and vocational assessment are the same as those for other types of assessment in principle. In practice, there are differences however. The following outlines the context of these qualifications to help elucidate why reliability is sometimes seen as less important in this field than in general qualifications.
While there are significant differences in the tradition of assessment in general education and in vocational education and training (VET), the newer emphasis in general education on outcomes and standards-referenced reporting has brought the two sectors closer together. As with VET, assessment in many areas of general education covers 'how to' (or procedural knowledge) as well as knowledge 'about' (or propositional knowledge). A common criticism is that general education has often over-valued the latter at the expense of the former, while VET is seen to have the reverse tendency.

VET has developed a much greater stress on the meaning of the individual standard of performance in the context of employment; that is, can the person do tasks well enough to meet employer needs? In VET while there has been an emphasis on common descriptors of outcomes (competencies) there has not been the emphasis on common assessment tasks that has characterized the external testing and public examination culture in general education. This creates difficulties in addressing the reliability agenda with psychometric and other technical methods. The following description of the approaches to vocational assessment has drawn on the report produced as part of the Ofqual Reliability Research Programme by Harth and van Rijn (2010), supplemented by reference to other publications.

National Vocational Qualifications and the Qualifications and Credit Framework

Vocational assessment occurs in the context of the *Qualifications and Credit Framework* (QCF). The QCF is the framework for the recognition and accreditation of *National Vocational Qualifications* (NVQs) in England, Wales and Northern Ireland. Under this framework (Ofqual, 2008, p.26), assessments are required to:

- be valid in relation to the learning outcomes against the stated assessment criteria
- produce sufficient evidence from learners to enable reliable and consistent judgments to be made about achievement of all the learning outcomes against the stated assessment criteria
- be manageable and cost effective
- be accessible

Since their introduction competence–based National Vocational Qualifications have been used primarily for employment purposes such as:

- confirmation of occupational competence
- licence to practice
- monitoring learner progression
- providing feedback to candidates for future improvement
- evaluating the effectiveness of assessor performance

Such vocational qualifications are outcomes-based, with no prescribed learning programmes and involve training in vocational areas ranging from construction, engineering, health and social care, service industries to business administration and management. The competence-based assessment process which is an essential feature of these qualifications consists of specification of standards, specification of opportunities to collect sufficient evidence, assessor judgments, learner feedback and quality assurance.

The concept of competencies intrinsic to modern approaches to vocational education is somewhat fuzzy and pragmatically developed to focus on knowledge and skills related to specific industry needs. Competencies can be considered 'as complex ability constructs that are context-specific, trainable, and closely related to real life' (Koeppen et al, 2008, p.61). In the context of the National Vocational Qualifications competence is about persons who have the ability to carry out activities to the required national occupational standards.

QCF units have assessment criteria which specify the content and range expected of the learner to achieve the learning outcome at the level of the unit. Evidence from tasks demonstrating achievement of these criteria is recorded. Achievement of these criteria indicates satisfactory performance of the function covered by the national occupational standard.

An important issue in considering assessment in VET is the type of data and methods of collection used which varies across the range of VET areas. A diverse range of procedures for accumulating evidence is used in VET (see Table 2 in Harth and van Rijn, 2010.) While most of the approaches described occur within different fields in general education, what characterizes their use in VET contexts is that different approaches may be used to provide acceptable evidence for the same outcome and are rarely standardized in terms of common tasks.

Competency based assessment (CBA)

Before the reform of vocational training, the traditional training emphasis was on the curriculum with a focus on the outcomes of the learning session. Standards of assessment were derived from job descriptions and training objectives designed to achieve the requirements of the industry. Assessment of competence involved carrying out a test in the training location to find out if the outcomes had been mastered. Such a test may have been written, practical or oral. Such tests often involved tasks which were indirect proxies for workplace events.

The more recent shift to *competency based assessment* (CBA) involves the assessment of evidence to judge a person's current abilities against a given set of standards or competencies set by an industry or enterprise to meet industry or enterprise needs. The assessment is designed to measure what a person can do in the workplace. It is the emphasis on knowledge-

as-action which has characterized this new approach. There is less interest in the process that produces the outcome than in the achievement of the outcome.

Competency standards involve specifying competencies related to the needs of the workplace as defined by employers. CBA involves the assessment of skills and knowledge to specific standards and geared to job needs. The result is not bound by time or curriculum, but is determined by demonstrated job skills. In a sense the alignment between curriculum and formative assessment (assessment for learning) and summative assessment which is such a big issue in general education (Biggs, 1996) is not an issue in VET, where the relationship can be characterized as assessment *as* learning, or learning *through* assessment.

The competencies assessed embrace the ability to perform a whole range of activities in a specific occupational area including transferring skills and knowledge to new situations and managing a variety of tasks within a job. Competency standards are set at a level for satisfactory performance required by industry. The emphasis is on getting all students to the same standard acceptable for practice. While the threshold for the performance standard is set in such a way to encourage skilled performance, typically the emphasis is on attaining the minimum necessary for being able to get employed in the industry. This has led to a perception that performance beyond the standard is not as valued in VET as it should be.

Clearly there are some contexts in which it does not make much sense in differentiating performance into proficiency levels: a person can either start up a machine or shut it down or they cannot. Other skill tasks may involve opportunities for degrees of performance beyond a threshold standard and these differences may be valuable to capture for employers who wanted higher levels of productivity or efficiency in the people they are hiring. Many people in the VET area do not like the idea of graded competencies, although when possible many training providers make such reports about the students they have assessed.

The claims for CBA are that it is both valid and reliable (Rutherford, 1995). Validity comes from the fact that assessment occurs on-the-job or as near to it as possible. Hence the performances assessed are essentially the skills needed to be demonstrated in an everyday work environment. Samples of real work performance remove any ambiguity about requirements. Such directness of the assessment is very appealing to employers who see it delivering assurance about the abilities of their workers.

The claim often made about CBA is that because CBA competency standards are written so that there is no ambiguity about their meaning assessors can make consistent and reliable judgments (Rutherford, 1995). This strong claim has been contested by Wolf (1995), who showed that even tightly written specifications of criteria are capable of multiple interpretations. The response to such criticism has involved a major push toward the development of a community of practice with a strong emphasis on the training and monitoring of assessors.

Quality Assurance and Verification of Assessments

The approach in CBA has been to quality assure (QA) the assessment process and to assume consistency of judgment follows from the training of industry assessors and their formal accreditation. Awarding organisations need to ensure 'the accuracy and consistency of standards in the assessment of units, across units and over time' (Ofqual, 2008, paragraph 5.6c). Thus awarding organisations establish procedures which 'require sufficient evidence from learners to enable reliable and consistent judgments to be made about the achievement of all the learning outcomes against the stated assessment criteria' (Ofqual, 2008, p.26).

Candidates are assigned one or several work-based or peripatetic assessors who are responsible for formally judging the evidence obtained from the candidate against the required assessment standards. These assessors are required to select assessment methods appropriate to the prescribed quality criteria, help candidates identify opportunities to demonstrate their competence, or produce evidence, especially when it is not possible to generate it as part of the normal work practice and when supplementary sources of evidence need to be generated (Fletcher, 1991). They also have to achieve relevant assessor qualifications for their role in order to be able to operate independently, participate in standardisation events and demonstrate that they are continuously updating their occupational competence and assessment skills. Assessors emphasise the developmental role of the vocational assessment which helps the learner to compare their performance to the standard required in their job roles.

Assessors and internal verifiers (IVs) use evidence collected over repeated occasions to decide whether the assessment criteria for a particular unit indicates that the candidate is:

- competent: the collected evidence meets the assessment criteria
- not yet competent: the candidate has not yet demonstrated all of the assessment requirements, either based on sufficient evidence or due to insufficiency of evidence where the candidate may not have had enough opportunities to perform the tasks.

Given that in many contexts students can repeat activities until competence is demonstrated, the main emphasis in the internal verifivation process is on the sufficiency of the evidence on which the assessor deems the student to be competent. The internal verification process is a key element of the QA process for providers of vocational qualifications. With respect to NVQs for example the *Joint Awarding Body Guidance on Internal Verification of NVQs* states that verification involves sampling assessments, monitoring assessment practice and standardising assessment judgments. The second element of the QA process involves external verification. The external verifier (EV) samples the evidence and decisions of the provider according to an agreed sampling frame. These verification processes are expected to ensure moderation and to produce consistency of judgments across providers of the qualification. The emphasis in the verification process is on quality assurance and process improvement. The purpose of the verification system is to ensure the consistency of decisions, through a complex set of relationships between assessors, internal verifiers (IV) and external verifiers (EV) and Awarding Organisations.

With the range and complexity of vocational education this emphasis on the front end of assessment procedure is understandable but there are still issues about the outcomes. The

debate about the claim that CBA avoids the reliability issues that arise more generally in educational assessment needs to be resolved by reference to empirical evidence. However in his useful review of recent vocational research in the UK, Johnson (2006, p.37) pointed out that 'there are significant gaps in recent reliability research … and it appears that we need to go back some time to find empirical work addressing such issues'.

Reliability in VET

Benett (1993) and Kane (2004) have argued that the fundamental principles of classical test theory may be applied to the more qualitative assessment of competencies, but Baartman et al. (2006) have challenged this claim. They pointed out that, traditionally, reliability is determined by consistency of measurement over repeated occasions given fixed raters, or in terms of internal consistency measures. They argue that consistency of repeated measures over time does not fit well with the developmental emphasis on shaping performance to converge on the required competency standard. With respect to internal consistency measures, they are less appropriate in VET where, instead of multiple test items, whole task performance is commonly used.

They conclude that the traditional statistical procedures used with objective tests to establish reliability are inappropriate for competency assessment and work-placed learning. Citing Gipps (1994) they say that 'We should abandon the idea that assessment is an exact science in which a 'true score' can be found' (Baartman, et al, 2006, p.156). While their alternative approach is to develop a framework of quality criteria for competency assessed programs, one criterion is *reproducibility of decisions.* By this criterion they focus on the decision made on the basis of evidence accumulated in a competency assessment program (CAP).

Given the assessment tradition developed within the VET sector it is not surprising that there is not a large literature directly relevant to the reliability of vocational assessment outcomes. Assessors in different contexts (candidates, centres) use different sources of evidence yet the same criteria to classify candidates. It is thus necessary for the decisions to be consistent in view of varying evidence when compared to a fixed outcome. Consistency or replication in this context would be that if the candidate were to be judged again, the same judgment should be made based on the evidence provided.

In their review of the literature on vocational qualifications Harth and van Rijn (2010, p.23) observe:

> Although no examples were found in the literature of methods suitable to estimate the classification reliability for assessments with varying number or type of tasks (a candidate may perform different tasks to achieve a fixed outcome), a number of studies of NVQs have used classical test theory methods, analysis of variance or qualitative methods of investigation such as interviews, questionnaire and field studies (see Wilmut, Wood and Murphy, 1996; Greatorex, 2000; Johnson, 2006 for reviews). This work has been important in understanding the challenges imposed by work-based assessment in the context presented by these qualifications, but has provided limited evidence for estimating the consistency of these decisions. Limited access to assessment data, logistical issues and time constraints have restricted advances in educational measurement theory of assessor decisions for these

qualification types.

The few recent articles based on UK practice in VET assessment suggest the need for further study. Greatorex and Shannon (2003) report that the Joint Awarding Body guidance project found that in fact little standardisation was carried out. They suggest that the reason that centres did not standardise assessment decisions is 'that they think that they are standardising them by standardising the assessment process and operationalising the internal verification system. Indeed centres are unaware that they are not standardising assessment decisions' (p.5). In their study of assessors of Retail Operations Level 2, Greatorex and Shannon found that assessors believed that following the same procedure would ensure consistent results although in practice their judgments were not always consistent. This is a 'due process' approach to reliability.

Another study published by Greatorex (2005) addressed the relevance of different types of evidence to consistency of competence judgments and concluded that there was an effect on assessor judgments. She concluded that 'further research is required before we can say precisely which types of evidence affect assessor judgments and in which circumstances' (p.162). Clearly more research would be welcome, but it must be taken as an act of faith that one could ever say 'precisely' which type and in which circumstances.

Options for Addressing Reliability in VET

Regulators of VET systems require providers to make assessments that are reliable (OfQual, 2008, p.26). However the practice of VET assessment raises issues about how the requirement of reliable assessment is to be met when common assessment tasks are not a consistent requirement. The large number of assessment decisions involved in some vocational qualifications and the flexibility afforded in the type of evidence considered makes it very difficult to address consistency.

Recommendation 10 **Greater consistency and control of assessment formats in work-place assessments should be required by Ofqual for new assessments, unless a rationale can be produced by awarding organisations for the validity and reliability of less well controlled assessments.**

Clearly moderation of assessor decisions is part of the current assessment process. However it is based on a sampling methodology and improvement culture rather than as an audit process with independent observations to enable conventional estimate of classification reliability. Nevertheless concerns about consistency of judgment need to be addressed. Awarding organisations could be required to provide data on consistency from their verification process.

The issue of how to audit assessor consistency has been the subject of discussion in many VET systems. One option considered in Australia is statistical moderation, but this would require change to practice:

> Although yet to be pursued at the national level within the VET Sector, statistical moderation could be used to ensure that RTO [Registered Training Organisation] based assessments are comparable throughout the nation, particularly if grades or marks are to be reported. However, to implement this moderation process, some form of a common assessment task(s) would need to be introduced at a national level in the VET sector (e.g. external exam or standardised assessment tools) to moderate the organisation-based assessments …The major benefit of statistical moderation is that it provides the strongest form of quality control over organisation-based assessments. It can also be less expensive to implement and maintain (if paper-based) than external moderation processes. It would however require the introduction of some form of common assessment task(s) at the national level. If the common assessment task was paper-based (as has been typically implemented in other educational sectors due to reduced costs associated with the implementation and scoring procedures), then any adjustments to candidate results would be limited to estimates of candidates' cognitive skills (i.e. knowledge and understanding); and therefore may have limited face and content validity within the VET sector.
>
> (National Quality Council, 2009, p.13-14)

This option has not received strong support from the VET sector in Australia as an appropriate solution to ensuring increased consistency of assessment. Currently their process is very similar to the quality assurance and external verification sampling model used in the UK.

Standard-setting reliability

Processes for setting standards are outlined by Ofqual's Code of Practice for many of the qualifications that they regulate, including the general qualifications (Ofqual, 2010). Despite specifying the process, this is not a mechanised system and value judgments are inherent in it (Baird, Cresswell and Newton, 2000; Baird 2007). Awarding body officials and senior examiners have to decide what weight they will give the different pieces of information that is before them regarding the quality of students' performances and the difficulty of the tasks they were set. As such, there is a question about how reliable the standard-setting judgments are. These judgments are generally made by panels of senior examiners and are largely accepted by Awarding Bodies' Accountable Officers. A different panel of examiners might make different decisions. Indeed, faced with slightly different evidence (for example a different sample of students' work on the same marks), the same panel might come to a different conclusion. Further, Awarding Bodies' procedures vary within the confines of the Code of Practice, so different types of information are used in the process and are presented at different times. These factors could also influence the reliability of the decisions. Baird and Dhillon (2005) investigated differences in procedures across the three English Awarding Bodies (AQA, Edexcel and OCR). Details of the procedures in which examiners scrutinised the students' work differed, such as whether examiners recorded their views about work on a particular mark (across a sample of scripts) or for an individual piece of work and whether the committee as a whole or individual examiners recorded decisions regarding the zone within which the grade boundary should lie. Five A level and five GCSE subjects were included in the study. Significant differences were found between two of the Awarding Bodies in terms of the proportion of consistent judgments between examiners (p.17). However, this does not tell us whether the overall reliability of the standard-setting would have been significantly different, as examiners' judgments of the students' work is only one part of the process. Furthermore, it is the collective judgment of the panel that counts. So what do we know about the reliability of standard setting for qualifications in England?

Research on the programme

Bramley and Dhawan (2010) simulated the impact of unreliability of standard-setting upon a small selection of GCSE, AS and A level examinations; looking specifically at the proportion of students who would be graded differently if the marks required for a grade were altered by one or more marks on the component question papers. Changing the grade boundaries by one mark had little impact upon students being awarded a grade A in AS business studies, AS chemistry and A level chemistry, but had a more dramatic impact upon the proportion of students who would have been awarded a grade A or grade C in GCSE psychology. The AS and A-level examinations were modular and so any particular question paper's grade boundaries had less of an effect upon the overall proportion of student attaining the grade. This research tells us about the likely impact of a small degree of unreliability in the process of standard-setting, but what we still do not know is the likely extent of unreliability of that process.

Research on replications of the standard-setting process

Only one study has been reported in which the standard-setting process has been replicated. Jones (2003, p.15) reports on a presentation by Michelle Meadows, in which she described the

replication of the standard-setting process using two separate senior examiner panels to set grade boundaries for a Biology and Human Biology paper. The grade-boundaries set were identically to the operational grade boundary decisions for the Biology paper, but were more severe for the Human Biology paper. This study only replicated the standard-setting process for individual question papers, which leaves out the parts of the process in which consideration is given to the outcomes for the subject as a whole and any comparisons across subjects that might be conducted either by the standard setting panel or the Accountable Officer. As such, we have slight evidence on this topic currently.

Research on consistency of examiners' judgments

Internationally, there have been studies of the consistency of individual examiners' judgments with each other in different types of standard-setting processes (Table 8). The conclusions from these studies vary and what counts as acceptable levels of consistency is not necessarily generalisable across context. Furthermore, we do not know the extent to which the findings of these studies will generalise to the kind of task that standard-setters conduct in England, which is specified by the Code of Practice (Ofqual, 2010) and does not entirely conform with the methodologies documented (see Cizek and Bunch, 2007) in the (often US-based) research literature.

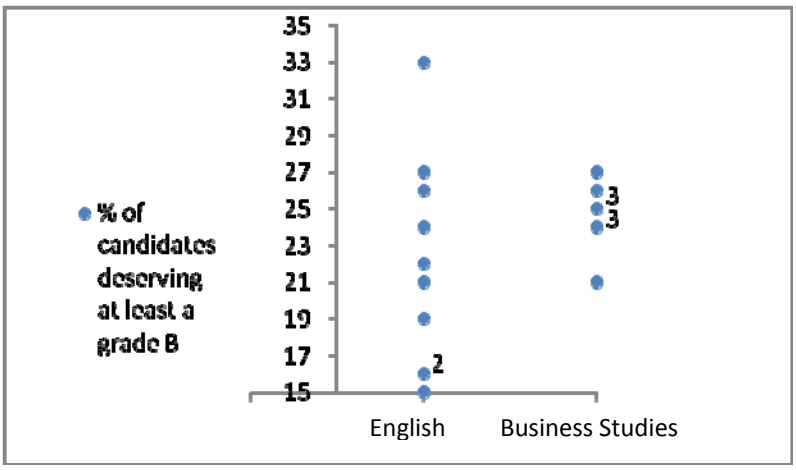
Table 8 Examples of international research on consistency of examiner judgments

Authors	**Standard-setting process**	**Test**	**Methodology**	**Brief summary of conclusions**
Jaeger (1989)	Nedelsky, Ebel, Contrasting Groups, Angoff, Borderline Groups, Jaeger	Various	Summary of literature	"… there is little consistency in … different standard methods" p.500
Verhoeven *et al* (1999)	Angoff	Medical progress test	Generalisability theory	Acceptable levels of error using 10 (recently qualified) judges on a 250 item test
MacCann and Stanley (2004)	Angoff and equi-percentile	New South Wales Year 10 School certificate	Central limit theorum, standard error	Reasonable agreement between methods in estimating consistency.
Plake, Impara and Irwin (2000)	Angoff	Certification in financial management	Descriptive statistics (means and standard deviations)	Consistency within and across panels and years

Authors	Standard-setting process	Test	Methodology	Brief summary of conclusions
Raymond and Reid (2001)	Review of the literature on several approaches (Angoff, Nedelsky, Jaeger, test-centred methods)	Various	Various	Most studies reviewed suggested 10-15 participants needed to achieve a dependable average judgment (dependability in the .8 range) (p.139)
Hurtz, G.M. and Auerbach, M.A. (2003)	Angoff	Various	Meta-analysis	Highest degree of consensus amongst judges found in conditions using a common standards definition and allowing discussion.
Wayne *et al* (2007)	Angoff and Hofstee	Clinical skills examination	Intraclass correlation	Both methods produced reliable data. Baseline data influenced judges' decisions.

Studies of the inter-examiner consistency in judging grade boundaries have also been conducted on English examinations under experimental conditions. Most of these studies were designed to investigate a particular feature of the process (Table 9), but data from them has also shed light on inter-examiner consistency. Examiners' grade boundary judgments for an English literature question paper had a standard deviation of 1.2 marks and for a psychology question paper it was 5.6 (Baird, 2000, Table 4). In Baird and Scharaschkin (2002), we see that the proportion of students who would have been awarded a grade B in A level English varied between 32% and 15% using individual examiners' holistic qualitative judgments alone (Figure 4). In A level business studies, the range was smaller, at 19-27 percent.

Figure 4 Differences in outcomes with different examiners' holistic judgments at grade B for two A-level subjects



Adapted from Baird and Scharaschkin, 2002, Figures 4 and 5
Figures on the graph indicate multiple examiners at that data point

Two studies have looked at the accuracy with which examiners can grade students' work within a small range of marks, as this is the task that the standard setting panels are faced with. Both studies found that examiners were not very accurate (Baird and Dhillon, 2005; Forster, 2005).

Research on the impact of features of the standard-setting process

Several studies have been conducted upon the impact of specific aspects of the English general qualifications' standard-setting process, but these do not tell us about the reliability of the process as a whole (Table 9).

Table 9 Some studies on aspects of the English general qualification standard-setting process

Authors	Process investigated	Brief summary of findings
Baird (2000)	Grade exemplar script – From correct or different grade	No effect on judgments in one subject; correct exemplar produced more correct judgments in another subject
Baird and Dhillon (2005)	Order of script scrutiny – mark order or random order	Mark order produced more undecided judgments
	Accuracy of judgments – grading judgments within a seven point mark range	Accuracy low within the small mark range typically used in operational standard setting judgments
Baird and Scharaschkin (2002)	Holistic judgment – Question paper or qualification level	Tunnel vision resulted from question paper judgments (more severe than qualification level)
Forster (2005)	Accuracy of judgments – Rasch analysis of Thurstone-pairs judgments within a small rang e of marks	Accuracy low for scripts 4 marks apart
Scharaschkin and Baird (2000)	Consistency of students' work – balanced or unbalanced performances in scripts	Consistency of student performance affected grading judgments.

Future research

Many questions remain unanswered regarding the reliability of standard setting processes. Although it would be logistically and politically difficult to conduct, a study investigating the impact of different Awarding Body procedures upon the reliability of standard setting would add a great deal to our knowledge. Replications of the standard-setting process such as the study conducted by Meadows (reported by Jones, 2003) would also be informative. With the findings on the reliability of examiner judgments being mixed (at best), research on the effect of statistics in the process and the impact that might have upon improving the reliability of judgments is also warranted.

Reliability of standard-setting is part of the wider issue of comparability of qualifications. This topic is also regulated by Ofqual and its predecessor sponsored a book on methodologies for conducting comparability studies (Newton, et al, 2007). Future research on the reliability of standard-setting could be pursued under the umbrella of comparability research.

Regulation

Ofqual has a remit to regulate standards and the consistency of standard setting practices and outcomes across years, awarding bodies and subjects is already part of its monitoring programme for general qualifications. Whilst the larger Awarding Bodies publish information on how standards are set, this is not a common feature across all organisations.

Recommendation 11 **Ofqual should require all examining bodies to document and publish their standard setting practices, so that regulation of standard setting reliability is more transparent in all sectors.**

Concluding remarks

One issue that has not been addressed in the programme or the preceding report is the matter of the reporting scale. The number of grades, levels and scores available affect the reliability of an assessment. We would recommend that the Department for Education conduct some work on the suitability of the reporting scales for educational assessments in England. At its most basic level, such research could investigate whether our assessments are reported in the correct number of grades. If we have highly unreliable assessments, they might not support the reporting of so many grades. Alternatively, we might have assessments that are so reliable that we could report more grades, which could be highly desirable for some assessment users, such as university entrance selectors.

The programme has met its remit in terms of production of evidence of reliability and interpretation of that evidence. Yet, the picture that has emerged is complex, with differing methodologies available and no principled way of choosing between some of them. As such, this report has made recommendations for minimum requirements, with the intention that where these are not suitable approaches, the awarding organisation will put forward an argument for an alternative, more suitable, approach or measure of reliability.

Despite concluding that the programme has met its remit, there is still a lot unknown about what levels of reliability should be deemed acceptable for educational assessments. This state of knowledge is recognised in testing standards produced by other bodies. Nonetheless, we envisage that by collecting a body of evidence regarding reliability of different assessment types, Ofqual should be able to produce an empirically-based set of standards within five years. Collection of information about the curriculum and assessment format alongside particular reliability indices will be essential to the production of contextualised standards.

References

ACME (2005) *Assessment in 14-19 Mathematics*. Advisory Committee on Mathematics Education, London: The Royal Society.

Baartman, L.K.J., Bastiaens, T.J., Kirschner, P.A., and van der Vieuten, C.P.M. (2006). The wheel of competency assessment: presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation*, 32, 153-170.

Baird, J. (2000) Are examination standards all in the head? Experiments with examiners' judgments of standards in A level examinations. *Research in Education*, 64, 91-100.

Baird, J. (2007) Alternative conceptions of comparability. Chapter 4 in Newton, P., Baird, J., Golstein, H., Patrick, H. and Tymms, P. (Editors) *Techniques for monitoring the comparability of examination standards*. QCA.

Baird, J. Cresswell, M.J. and Newton, P. (2000) Would the *real* gold standard please step forward? *Research Papers in Education*, 15, 213 – 229.

Baird, J. and Dhillon, D. (2005) *Qualitative expert judgments on examination standards: valid but inexact.* Assessment and Qualifications Alliance Internal Report, RPA-05-JB-RP-229.

Baird, J. and Scharaschkin, A. (2002) Is the whole worth more than the sum of the parts? Studies of examiners' grading of individual papers and candidates' whole A-level examination performances. *Educational Studies,* 28, 143 – 162.

Benett, Y. (1993). The validity and reliability of assessments and self-assessments of work-based learning. *Assessment & Evaluation in Higher Education*, 18, 83-95.

Bentler, P. M. (2009) Alpha, dimension-free, and model-based internal consistency reliability . *Psychometrika,* 74, 137-143.

Bentler, P., and Woodward, J. (1980). Inequality among lower bounds to reliability: With applications to test contruction and factor analysis. *Psychometrika,* 45, 249-267.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32, 347-364.

Black, P., Harrison, C., Osborne, J. and Duschl, R. (2004) *Assessment of Science Learning 14-19*. London: Royal Society. ISBN 0 85403 598 2. Available on: www.royalsoc.ac.uk/education.

Black, P., Harrison, C., Hodgen, J., Marshall, M. and Serret, N. (2010) Validity in teachers' summative assessments. *Assessment in Education,* 17(2), 215-232.

Borsboom, D. (2006) The attack of the psychometricians. *Psychometrika*, 71 (3), 425 – 400.

Bramley,T. and Dhawan, V. (2011). Estimates of reliability of qualifications. Report produced for Ofqual's Reliability Programme.

Brennan, R L (1983). *Elements of Generalizability Theory*. Iowa City: ACT Publications.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322.

Chamberlain, S. (2010) Public perceptions of reliability report. Ofqual Reliability Programme Report.

Cizek, G. and Bunch, M.B. (2007) *Standard Setting. A guide to establishing and evaluating peformance standards on tests.* Sage: California.

Cronbach, L. (1988) Five perspectives on validity argument. In Wainer, H. and Braun, H. (Eds.) *Test validity* (3 – 17) Lawrence Erlbaum: New Jersey.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* 16, 297-334.

Crooks, T.J., Kane, M.Y. and Cohen, A.S. (1996) Threats to the Valid Use of Assessments. *Assessment in Education,* 3(3), 265-285.

Cunningham, G. K. (1986). *Educational and psychological measurement*. New York: Macmillan.

Dimitrov, D.M. (2002) Reliability: arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement*, 62, 783 – 801.

E.U. Council of Ministers (2010). *Assessment of key competences: Draft Background Paper for the Belgian Presidency meeting for Directors-General for school education.* Brussels: E.U.

Evers, A., Lucassen, W., Meijer, R. and Sijtsma, K. (2009). *COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie)* [COTAN assessment sytem for the quality of tests (revised version)]. Amsterdam: NIP.

Fan, X. and Thompson, B. (2001) Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, 61, 517 – 531.

Fletcher S (1991). *NVQs, Standards and Competence*. Kogan Page: London.

Forster, M. (2005) *Can examiners successfully distinguish between scripts that vary by only a small range of marks?* Unpublished internal paper, Oxford, Cambridge and RSA.

Graham, J M (2006). Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability. *Educational and Psychological Measurement*, 66, 930-44.

Greatorex J (2000). What research can an awarding body carry out about NVQs? A paper presented at the British Research Association Conference, University of Cardiff, September.

Greatorex, J. and Shannon, M. (2003). How can NVQ Assessor's judgements be standardised? Paper presented at the British Educational Research Association Conference, 11-13 September, Heriot-Watt University, Edinburgh.

Greatorex, J. (2005). Assessing the evidence: different types of NVQ evidence and their impact on reliability and fairness. Journal of Vocational Education and Training, 57, No. 2, 149-164.

Green, S. B. and Yang, Y. (2009). Commentary on coefficient alpha. *Psychometrika,* 74, 121-135.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley & Sons, Inc.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika,* 10, 255-282.

Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research papers in education*, 20 (3), 245-270.

Harth, H. and van Rijn, P. (2010). Reliability issues in competence-based assessment: concepts and estimates. April. Ofqual Reliability Programme Report.

He, Q. (2009) Estimating the reliability of composite scores. Ofqual Reliability Programme Report.

He, Q., Hayes, M. and Wiliam, D. (2011) Classification accuracy in results from Key Stage 2 National Curriculum tests. Ofqual report.

He, Q., Opposs, D. and Boyle, A. (2010) A quantitative investigation into public perceptions of reliability in examination results in England. Ofqual Reliability Programme Report.

House of Commons (2008) *Testing and* Assessment. Children, Schools and Families Committee Report. Third Report on Session 2007-8.

House of Commons (2009) *Apprenticeships, Skills, Children and Learning Bill*. Bill 55. 4 February.

Hurtz, G.M. and Auerbach, M. A. (2003) A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement,* August, 63, 584-601.

Hutchison, D. and Benton, T. (2009). *Parallel universes and parallel measures: estimating the reliability of test results*. Report Ofqual/10/4709. Coventry: Ofqual.

Hutchison, D. & Schagen, I. (1994). *Reliability of Adaptive National Curriculum Assessment: Report of Project* Slough: NFER.

International Test Commission (2000) *International Guidelines for Test Use. Version 2000.* http://www.psychtesting.org.uk/download$.cfm?file_uuid=648F1A9C-CF1C-D577-94EB-E0B28414B0F4&siteName=ptc

Ipsos MORI (2009) *Public perceptions of reliability in examinations*. Available online at: http://www.ofqual.gov.uk/files/2009-05-14_public_perceptions_of_reliability.pdf.

Jaeger, R.M. (1989). Certification of student competence. Chapter 14 in Linn, R. (Editor) *Educational Measurement.* 3rd Edition. 485 – 515. National Council on Measurement in Education. American Council on Education. Macmillan Publishing Company, New York.

Johnson, S. (2011). *A review of the literature on teacher assessment reliability.* Report prepared for the OFQUAL Technical Advisory Committee.

Johnson, S. and Johnson, R. (2011). Conceptualising and interpreting reliability. Report Ofqual/10/4706. Coventry: Ofqual.

Johnson, M (2006). A review of vocational research in the UK 2002-2006: Measurement and accessibility issues. International Journal of Training and Research, 4, no.2, 48-71.

Johnson, S. and Munro, l. (2008). Teacher judgments and test results: should teachers and tests agree? *Paper presented at the annual conference of the Association for Educational Assessment – Europe, Hissar, Bulgaria, November 2008.*

Joint Awarding Bodies (2002). Joint Awarding Bodies Guidance on Verification of NVQs. London: Department for Education and Skills.

Joint Council for Qualifications (2010). *GCSE specifications and Principal Learning Units within Diploma Qualifications: Instructions for conducting controlled assessments, 1 September 2010 to 31 August 2011.* London: JCQ

Jones, B. (2003). *Report of the JCGQ Research Seminar on issues related to comparability of standards*. Assessment and Qualifications Alliance Internal Report, RPA-04-BEJ-RC-264.

Kane, M.T. (2004). Certification testing as an illustration of argument-based validation. Measurement: Interdisciplinary Research and perspectives, 2, 135-170.

Klenowski, V. (2009). Editorial – Assessment for Learning revisited: an Asia-Pacific perspective. *Assessment in Education* vol.16 (3), 263-268.

Koeppen, K., Hartig, J., Klieme, E. and Leutner, D. (2008). Current issues in competence modeling and assessment. Zeitschrift für Psychologie, 216 (2), 61-73.

Koretz, D. (1998). Large Scale Portfolio Assessments in the US: evidence pertaining to the quality of measurement. *Assessment in Education,* 5(3) 309-334

Kuder, G.F., and Richardson, M.W. (1937). The theory of estimation of test reliability. *Psychometrika,* 2, 151-160.

Linacre, J. M., and Wright, B. D. (2001). *A User's Guide to Winsteps*. Chicago: MESA Press.

Lindley, P., Bartram, D. and Kennedy, N. (2008). *EPPA Review Model for the Description and Evaluation of Psychological Tests. Test Review Form and Notes for Reviewers.* Version 3.42. http://www.psychtesting.org.uk/download$.cfm?file_uuid=696852C1-985D-B6B0-7E55-29D144984AF3&siteName=ptc

Leckie, G., and Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 172, 4, 835 – 851.

Lord, F.M. (1984). Standard error of measurement at different ability levels. *Journal of Educational Measurement*, 21, 239 – 243.

Lumsden, J. (1976). Test theory. In Rosenzweigh, M.R. and Porter, L.W. (Editors) *Annual Review of Psychology* (Volume 27), Palo Alto, Calififorniа: Annual Reviews.

MacCann, Robert G. & Gordon Stanley (2004). Estimating the standard error of the judging in a modified-angoff standards setting procedure. *Practical Assessment, Research & Evaluation*, 9(5). Retrieved August 9, 2010 from http://PAREonline.net/getvn.asp?v=9&n=5 .

MacCann, R.G. and Stanley, G. (2010). Classification consistency when scores are converted to grades: examination marks versus moderated school assessments. *Assessment in Education.* 17 (3), 255-272.

McDonald, R.P. (1999). *Test theory: A unified treatment.* Hillsdale: Erlbaum.

Mansell, W., James, M. and the Assessment Reform Group on behalf of TLRP. (2009). *Assessment in Schools. Fit for Purpose?* London: Institute of Education.

Masters, G. N. and McBryde, B. (1994). An investigation of the comparability of teachers' assessment of student folios. Brisbane: Queensland Tertiary Entrance Procedures Authority.

Maughan, S., Styles, B., Lin, Y. and Kirkup, C. (2009). *Partial Estimates of Reliability: Parallel Form Reliability in the Key Stage 2 Science Tests*. NFER Report. Slough: NFER

National Quality Council (2009). A code of professional practice for validation and moderation. www.tvetaustralia.com.au.

Newton, P.E. (2009). The reliability of results from national curriculum testing in England. *Educational Research*, 51, 2, 181-212.

Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Editors).(2007). *Comparability of UK public examinations*. QCA book publication.

Ofqual(2010). *GCSE, GCE, principal learning and project code of practice.* http://www.ofqual.gov.uk/for-awarding-organisations/96-articles/247-codes-of-practice-2010

Ofqual (2008). Regulatory arrangements for the Qualifications and Credit Framework. http://www.ofqual.org.uk/files/Regulatory_arrangements_QCF_August08.pdf

Plake, B.S., Impara, J.C. and Irwin, P.M. (2000). Consistency of Angoff-Based Predictions of Item Performance: Evidence of Tehcnical Quality of Results from the Angoff Setting Method. *Journal of Educational Measurement*, 37, 4, 347 – 355.

Popham, J.W. (2007). Instructional Insensitivity of Tests: Accountability's Dire Drawback. *Phi Delta Kappan*, 89, 2, 146-150.

Porter, T.M. (1986). *The rise of statistical thinking 1820 – 1900.* Princeton University Press: New Jersey.

QCA (2006). *A review of GCSE coursework.* London: Qualifications and Curriculum Authority.

QCA (2007). *Controlled assessments*. London: Qualifications and Curriculum Authority.

QCDA (2009). Changes to GCSEs and the introduction of controlled assessment for GCSEs. London: Qualifications and Curriculum Development Agency.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria. (ISBN 3-9000051-07-0)

Raymond, M.R. and Reid, J.B. (2001). Who made thee a judge? Selecting and training participants for standard-setting. Chapter 5 in Cizek, G.J. (Editor), *Setting Performance Standards*, Lawrance Erlbaum Associates, New Jersey.

Reeves, D.J., Boyle, W.F. and Christie, T. (2001). The relationship between teacher assessments and pupil attainments in standard test tasks at keystage 2 1996-98. *British Educational Research Journal.* 27(2) 141-160.

Revelle, W. (2008). *Psych: Procedure for personality and psychological research* (R package version 1.0-51).

Revelle, W., Zinbarg, R.E. (2009). Coefficient alpha, beta, omega, and the GLB: comments on Sijtsma. *Psychometrika, 74,* 145-154.

Rutherford, P.D. (1995). Competency based assessment. Melbourne, Pitman.

Saal, F.E., Downey, R.G. and Lahey, M.A. (1980). Rating the ratings – assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413 – 428.

Scharaschkin, A., and Baird, J. (2000). The effects of consistency of performance on A level examiners' judgments of standards. *British Educational Research Journal*, 26, 343 – 357.

Schools Council (1980). *Focus on examinations.* Pamphlet 5. London, Schools Council.

Shavelson, R and Webb, N (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shapley, K.S. and Bush, M.J. (1999). Developing a valid and reliable Portfolio Assessment in the Primary Grades: Building on Practical Experience. *Applied Measurement in Education,* 12(2), 111-132.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefullness of Cronbach's alpha. *Psychometrika, 74,* 107-120.

Smith, G.A. (1978). *JMB experience of the moderation of internal assessments. OP38.* Manchester: Joint Matriculation Board.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295.

Stanley, G., MacCann, R., Gardner, J., Reynolds, L. and Wild, I. (2009). *Review of teacher assessment: what works best and issues for development.* Oxford University Centre for Educational Development; Report commissioned by the QCA. http://www.education.ox.ac.uk/assessment/uploaded/2009_03-Review_of_teacher_assessment-QCA.pdf

Taylor, M. (1992). *Reliability of marking of coursework for GCSE and GCE*. Guildford: AQA.

Ten Berge, J. M. F., and Sočan, G. (2004). The greatest lower bound to reliability of a test and the hypothesis of unidimensionality. *Psychometrika,* 69, 613-625.

Thompson, B. and Cook, C. (2002). Stability of the reliability of LibQual+TM scores. A reliability generalization meta-analysis study. *Educational and Psychological Measurement*, 62, 735 – 743.

Tymms, P.B. (1996). *The Value Added National Project: Second Primary Technical Report: an analysis of the 1991 Key Stage 1 assessment data inked to the 1992 KS2 data provided by Avon LEA* London: School Curriculum and assessment Authority.

van Lent, G., Watts, A. and Wools, S. (2010). *Towards a European Framework of Standards for Educational Assessment.* http://www.aea-europe.net/page-292.html

Verhelst, N.D. (1998). Estimating the reliability of a test from a single test administration (*Measurement and Research Department Report 98-2).* Arnhem, The Netherlands: Cito Institute for Educational Measurement.

Verhelst, N.D., Glas, C.A.W., and Verstralen, H.H.F.M. (1993). *OPLM: One Parameter logistic model. Computer program and manual*. Arnhem: Cito.

Verhoeven, Van der Steeg, Scherpbier, Muijtjens, Verwijnen and Van Der Vleuten (1999). Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. *Medical Education*, 33: 832–837. doi: 10.1046/j.1365-2923.1999.00487.x

Wayne, D.B., Barsuk, J.H., Cohen, E. and McGaghie, W.C. (2007). Do baseline data influence standard setting for a clinical skills examination. *American Medicine*, 82, 10, 105-108.

Wheadon, C. and Stockford, I. (2011). Classification Accuracy and Consistency in GCSE and A Level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009. Report produced for Ofqual's Reliability Programme.

Whetton, C. (2009). A brief history of a testing time: national curriculum assessment in England 1989-2008, *Educational Research*, 51(2): 137-159.

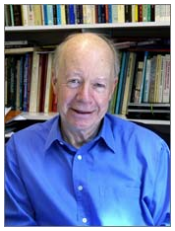
Wilmut J, Woods R and Murphy R (1996). A Review of Research into the Reliability of Examinations. A discussion paper prepared for the School Curriculum and Assessment Authority. http://www.nottingham.ac.uk/shared/shared_cdell/pdf-reports/relexam.pdf

Wolf, A. (1995). Competence-based assessment. Buckingham, UK: Open University Press.

Appendix A The Technical Advisory Group

Professor Jo-Anne Baird is Lead Editor of the academic journal Assessment in Education: Principles, Policy & Practice and Director of the University of Bristol's Centre for Assessment and Learning Studies. She was previously Head of Research at the Assessment and Qualifications Alliance, where she was responsible for managing the research programme and for the standard-setting process for AQA's examinations. Jo-Anne also worked as a Lecturer at London University's Institute of Education and has taught for the Open University and at A level. Jo-Anne is a Fellow of the Association of Educational Assessment-Europe. In 2007, Jo-Anne co-edited the book, *Techniques for comparing examination standards*, which was commissioned by the Qualifications and Curriculum Authority. Her research looks at examination standards and systemic issues in assessment quality.

Dr Anton Béguin is Director of the Measurement and Research Department at Cito, an international measurement and assessment organisation, where he has worked since 2001. He has been responsible for statistical and psychometric procedures used in the central examinations in the Netherlands. Dr Béguin is one of the project directors of a large scale longitudinal study in the Netherlands and has been involved in many research projects related to standard setting and equating. He has worked as consultant advising on measurement, testing and accountability with a range of national and international organisations. He began his career as a research assistant at University of Twente in 1995, where he also wrote his doctoral thesis. Dr Béguin has written many publications on examinations and measurement and is the Chair of the Methodology and Evaluation division of the Netherlands Educational Research Association (NERA), and Fellow of the Association of Educational Assessment-Europe.



Professor Paul Black is Emeritus Professor of Education at King's College London. He took his first degree in physics at the University of Manchester, followed by a PhD at Cambridge University. He spent 20 years a faculty member of the Department of Physics at the University of Birmingham and in 1976 he became Professor of Science Education and Director of the Centre for Science and Mathematics Education at Chelsea College in London. When Chelsea College merged with King's in 1985, he became the head of the King's Centre for Education Studies, King's College London. Professor Black was also Chair of the government's task group on assessment and testing in 1987-88 and Deputy Chairman of the National Curriculum Council from 1989 to 1991. He is an honorary life member and former president of the Association for Science Education (UK), and in 2004 he received the lifetime service award for Distinguished Contribution to Research in Science Education (US).

Alastair Pollitt set up the Cambridge Exam Research consultancy, which offers research and training on formal assessment, last year. Before this, he was made

the Director of Research and Evaluation at the University of Cambridge Local Examinations Syndicate (UCLES) in 1994, where he stayed until 2004. He began his career at the University of Aberdeen, where he graduated in chemistry with mathematics. He spent two years teaching science in the West Indies which sparked his interested in learning, and spurred him to take an MA in Education and Psychology at the University of Aberdeen. He then trained in psychometrics and statistics in Edinburgh. From 1980 to 1983 Mr Pollitt embarked on a programme of government-sponsored research to investigate the psychology of examining, which resulted in his 1985 book What Makes Exam Questions Difficult? He was director of the national project to monitor standards in English language amongst Scotland's schoolchildren in 1989. Mr Pollitt is currently working on a project with the University of London to explore awarding grades using holistic comparative judgement.

Professor Gordon Stanley is a Visiting Research Fellow at the University of Sydney. He was the first Director of the Oxford University Centre for Educational Assessment, as well as a research fellow of St Anne's College and Pearson Professor of Educational Assessment at the University of Oxford. Prior to taking up this position at Oxford, he spent 10 years heading a curriculum and public examinations authority in New South Wales, Australia. His career has involved teaching and research in assessment as well as holding statutory offices in education. In 1997 he was made Chair of the Australian National Board of Employment Education and Training and of the Higher Education Council. During the 1990s he became involved in quality assurance issues in education and was a member of the Committee for Quality Assurance in Higher Education. In his NBEET role he provided advice on quality issues to the Commonwealth Minister for education. Professor Stanley is an Emeritus Professor of Psychology at the University of Melbourne and Honorary Professor of Education at the University of Sydney.

Professor Dave Bartam was a member of the Technical Advisory Group in 2009. Dave is Research Director at Saville and Holdsworth Limited. He joined SHL in 1998, having been a Faculty Dean and Professor of Psychology at the University of Hull. He is past Chair of the British Psychological Society (BPS) Steering Committee on Test Standards and Chair of the EFPA Standing Committee on Tests and Testing. He is President of the International Association of Applied Psychology's Division 2 (Measurement and Assessment) and a past-President of the International Test Commission. He has received the BPS award for Distinguished Contribution to Professional Psychology and has been widely published in a range of areas relating to occupational assessment, especially in relation to computer-based testing.

Appendix B Contributors to the Reliability Programme seminar series

Fekhir Al-Naeme	Scottish Qualifications Authority (SQA)
Angus Alton	Ofqual
Cathy Barnes	Signiture
Andrew Boyle	City & Guilds
Jenny Bradshaw	National Foundation for Education research (NFER)
Tom Bramley	Cambridge Assessment
Charmain Campbell	City & Guilds
Susan Chamberlain	Assessment and Qualifications Alliance (AQA)
Chung-Pak Cheung	Oxford Cambridge & RSA Examinations (OCR)
Rose Clesham	Edexcel
Robert Coe	CEM Durham University
Mike Cresswell	Assessment and Qualifications Alliance (AQA)
Denver Davies	The Department for Children, Education, Lifelong Learning & Skills (DCELLS)
Vikas Dhawan	Cambridge Assessment
Barbara Donahue	Qualification & Curriculum Development Agency (QCDA)
Inga Fitzgerald	City & Guilds
Mike Forster	Oxford Cambridge & RSA Examinations (OCR)
Jeffrey Goodwin	Independent
Elizabeth Gray	Oxford Cambridge & RSA Examinations (OCR)
Silvia Green	Cambridge Assessment
Ducan Grimes	General teaching council for England (GTCE)
Helen Harth	City & Guilds
Malcom Hayes	Edexcel
Karen Hughes	Edexcel
Stephen Hills	City & Guilds
Dougal Hutchison	National Foundation for Education research (NFER)
Tina Isaacs	Institute of Education
Rod Johnson	Assessment Europe
Sandra Johnson	Assessment Europe
Mike Kingdon	Independent
Janne Karkkainen	Edexcel
Pamela Lamour	Council for the Curriculum, Examinations & Assessment (CCEA)
Louise Maycock	Qualification & Curriculum Development Agency (QCDA)
Michelle Meadows	Assessment and Qualifications Alliance (AQA)
Sarah Maughan	National Foundation for Education research (NFER)
Paul McAndrew	Education Development International (EDI)
Melissa McConaghy	Council for the Curriculum, Examinations & Assessment (CCEA)
Paul Newton	Cambridge Assessment
Roger Murphy	University of Nottingham
Tim Oates	Cambridge Assessment
Gareth Pierce	Welsh Joint Education Committee (WJEC)
Anne Pinot de Moira	Assessment and Qualifications Alliance (AQA)

Jeremy Pritchard	Edexcel
Jo Richards	Welsh Joint Education Committee (WJEC)
Rachel Roberts	City & Guilds
Colin Robinson	Independent
Jenny Scharf	Council for the Curriculum, Examinations & Assessment (CCEA)
Stuart Shaw	Cambridge International Education (CIE)
Vikki Smith	City & Guilds
Nick Sofroniou	Welsh Joint Education Committee (WJEC)
Ian Stockford	Assessment and Qualifications Alliance (AQA)
Steve Strand	Warwick University
Ben Styles	National Foundation for Education research (NFER)
Ezekiel Sweiry	Qualification & Curriculum Development Agency (QCDA)
Kath Thomas	Edexcel
Raymond Tongue	Welsh Joint Education Committee (WJEC)
Liz Twist	National Foundation for Education research (NFER)
Rob Van Krieken	Scottish Qualifications Authority (SQA)
Peter Van Rijn	CITO
Chris Wheadon	Assessment and Qualifications Alliance (AQA)
Dylan Wiliam	Institute of Education
Jan Winter	Bristol University
Alison Wood	Qualification & Curriculum Development Agency (QCDA)

Appendix C Reliability Programme Remit

1. Reliability, in educational assessment terms, can be defined as consistency. A high level of reliability means that broadly the same outcomes would arise were an assessment to be replicated. Given the general parameters and controls that govern the assessment process (including test/exam specification, administration conditions, approach to marking, standard setting methodology and so on), reliability concerns the impact of the factors that inevitably vary from one assessment to the next. These include:

    - the particular **occasion** (e.g. if assessed on another day, the student might have been less tired)
    - the particular **test** (e.g. if a different test/exam had been set, the student might not have been confused by the wording of an essay title)
    - the particular **marker** (e.g. if a different marker had been assigned, the student might have been marked down for using an unusual stylistic construction)
    - the particular **standard setting** panel (e.g. if a different team of people had been involved, different grade boundaries might have been set).

2. In England, there has been little systematic and sustained attempt to evaluate the reliability of results from national tests and examinations. The work that has been undertaken has been:
    - isolated (i.e. not part of routine monitoring)
    - partial (i.e. limited to certain sources of unreliability and to a small number of tests and examinations)
    - under-theorised (i.e. with little serious debate over the interpretation of evidence)
    - under-reported (i.e. not always published)
    - misunderstood by stakeholders, both inside and outside assessment agencies.

3. A substantial programme of research into reliability will help to improve this situation.

4. The project will consist of three strands:
    - generating evidence of reliability
    - interpreting evidence of reliability
    - developing a policy on reliability.

Strand 1: Generating evidence of reliability

Aim

5. The aim of strand 1 will be to generate robust evidence of the overall reliability of results from a number of major national tests and/or examinations, estimating the degree of consistency associated with different aspects of the assessment process.

Methodology

6. The precise methodology will be subject to discussion with assessment experts and agencies. Not all sources of inconsistency will necessarily be investigated, although there will be a particular focus on test-related and marker-related inconsistency. The primary focus of attention will be on reliability at the student level, although implications for reliability at the cohort level will also have to be considered given the widespread use of aggregate scores for comparative purposes at national, regional and local levels.

7. Comprehensive estimates of reliability will require experimental simulation as well as the analysis of data which arise as a natural by-product of testing and examining. For example, to estimate the consistency of performance across test/exam forms, it may be necessary to administer alternative versions to the same students. To estimate the consistency of marking across scripts, it may be necessary to have batches of scripts marked by multiple markers. Ideally, these variables will be manipulated within a single experimental design.

8. It is desirable that, over time, such analyses will be undertaken across a range of subjects, for a range of tests, examinations and qualifications and considering both externally assessed and internally assessed components. Reliability estimates inevitably differ across contexts, being sensitive to a range of factors, from the group of candidates entered to the design of the assessment process, so estimates for one instrument cannot necessarily be assumed to generalise to another. In the long term, this might imply the need for a monitoring programme, rather than occasional studies.

9. In the short term, it would be wise to begin by focusing on a limited number of tests and/or examinations. Even starting with a small sample – perhaps English and mathematics tests at key stage 2 – the project will be substantial, complex and costly, due to the large number of variables to be manipulated experimentally.

Strand 2: Interpreting evidence of reliability

Aims

10. The aims of strand 2 will be to stimulate, capture and synthesise technical debate on:
    i. the interpretation of evidence from reliability studies
    ii. the communication of results from reliability studies.

Methodology

11. The interpretation and communication of evidence from reliability studies is a highly complex challenge which will require collaboration between assessment experts, agency representatives and communications specialists. It is likely that this strand will tackle the two aims sequentially, with assessment experts and agency representatives debating the interpretation of evidence from reliability studies before being joined by communications specialists to discuss the communication of results.

12. It will be necessary to identify the comparators against which reliability evidence from England's test and examinations can be benchmarked. These might include alternative assessment models, i.e. different approaches to testing/examining or different approaches

to teacher assessment, as well as test and examination systems from other countries which operate a similar approach to England.

13. The debates will be undertaken during residential workshops, with participants being provided with working papers in advance. Outcomes from the workshops will be circulated for comment following the workshops, resulting in a series of published reports.

Strand 3: Developing a policy on reliability

Aims

14. The aims of strand 3 will be to:
    i. explore public understanding of, and attitudes towards, assessment inconsistency
    ii. stimulate national debate on the significance of the reliability evidence generated by the project
    iii. develop a policy position for Ofqual on reliability.

Methodology

15. Many myths are promoted (particularly within assessment circles) about what the public understand about assessment inconsistency, and how they will react to evidence of reliability, particularly when framed in terms of the percentage of students whose grades are likely to be incorrect. The reality is that we simply do not know what the public thinks and feels on this matter.

16. This research will engage with members of the public – students, parents, employers and so on – listening to their views and beliefs, using a series of surveys and focus groups.

17. The findings will be promoted more widely, through engagement with the national media and through the use of discussion documents on the Ofqual website. These debates and discussions will help inform an Ofqual policy position on reliability that will need to be developed. The policy is likely to include both how public and professional understanding of reliability can be improved, including the evidence that needs to be generated to inform this understanding, and a position with regards to how reliability affects the reporting of results.

Appendix D Reliability Programme Reports and Policies

Reports of the Reliability Programme can be found at:
http://www.ofqual.gov.uk/research-and-statistics/research-reports/92-articles/20-reliability

Baird, J., Beguin, A., Black, P., Pollitt, A. and Stanley, G. (2011) The Reliability Programme: Final report of the Technical Advisory Group.

Bradshaw, J. and Wheater, R. (2010) International survey of results reporting

Bramley,T. and Dhawan, V. (2011) Estimates of reliability of qualifications

Burslam, S. (2011) The Reliability Programme: Final report of the Policy Advisory Group

Chamberlain, S. (2010) Public perceptions of reliability

Boyle, A., Opposs, D. and Kinsella, A. (2009) No news is good news? Talking to the public about reliability of assessment

Harth, H. and Hemker, B. (2011) On the reliability of results in vocational assessments

He, Q. (2009) Estimating the reliability of composite scores

He, Q. (2010) Ofqual's reliability of results programme

He, Q., Opposs, D. and Boyle, A. (2010) A quantitative investigation into public perceptions of reliability in examination results in England

He, Q., Opposs, D. and Boyle, A. (2010) Public perceptions of unreliability in examination results in England: A new perspective

He, Q., Boyle, A. and Opposs, D. (2010) Investigating the reliability of results from national tests and public examinations in England

He, Q., Hayes, M. and Wiliam, D. (2011) Classification accuracy in results from KS2 National Curriculum tests

Hutchison, D. and Benton, T. (2010) Parallel universes and parallel measures

Johnson, S. and Johnson, R. (2009) Conceptualising and interpreting reliability

Johnson, S. (2011) A focus on teacher assessment reliability in GCSE and GCE

Johnson, S. and Johnson, R. (2011) Component reliability in GCSE and GCE

Maughan, S., He, Q., Styles, B and Lin, Y. (2010) Reliability of results from national Curriculum Assessment, A UK perspective

Maughan, S., Styles, B., Lin, Y. and Kirkup, C. (2010) Partial estimates of reliability

MORI, (2009) Public perceptions of reliability

Ofqual. (2009) The Reliability Programme technical seminar report – 7 October 2009

Opposs, D. and He, H. (2010) The reliability of results from National Curriculum assessments, public examinations and qualifications

Opposs, D. and He, H. (2011) The Reliability Programme: Final report

Phelps, R., Zenisky, A., Hamleton, R. and Sireci, S. (2010) On the reporting of measurement uncertainty and reliability for U.S. educational and licensure tests

Wheadon, C. and Stockford, I. (2011) Classification Accuracy and Consistency in GCSE and A Level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2011

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346

www.ofqual.gov.uk